
DATA
REVOLUTION
大数据价值实现
方法、技术与案例
范煜◎著
宏观和微观、人文和技术、启迪思想
和关注实用并举
既适合宏观层面的领导启迪思维，提出工作目标，又适合
微观层面的执行人员找到实现的方法和路径
清华大学出版社

数据革命

大数据价值实现方法、技术与案例

范煜　著

清华大学出版社
北　京

内 容 简 介

在信息技术革命之后，我们将迎来数据革命。在大数据的概念、性质和价值已得到政府和社会的认可之后，大家关注的是数据如何获取，以及有了数据以后如何挖掘数据的价值。仅适合特定行业、满足特定需求的技术不足以应对一场革命，大数据不但是超出计算机软硬件处理的能力，更是超出人类的认知能力。只有实现对数据的认知，利用数据辅助决策，才是适合不同行业数据价值实现的通用手段。本书描述了数据革命的起源、实现的思路、所用的技术和要达到的目标，针对当今社会热点描述了在数据时代的应对之策。

本书宏观和微观、人文和技术、启迪思想和关注实用并举，既适合宏观层面的领导启迪思维，提出工作目标，又适合微观层面的执行人员找到实现的方法和路径。本书介绍的理论和技术均可在智慧城市、智能制造领域实际使用。

本书适合政府、企业决策者和 CIO，及其他对大数据应用感兴趣的人阅读。

图书在版编目(CIP)数据

数据革命：大数据价值实现方法、技术与案例 / 范煜著. — 北京：清华大学出版社，2017
ISBN 978-7-302-46693-2

Ⅰ. ①数… Ⅱ. ①范… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 038741 号

责任编辑：刘 洋
封面设计：李召霞
版式设计：方加青
责任校对：宋玉莲
责任印制：王静怡

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-62770175 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经 销：全国新华书店
开 本：170mm×240mm **印 张：**15.75 **字 数：**231 千字
版 次：2017 年 5 月第 1 版 **印 次：**2017 年 5 月第 1 次印刷
印 数：1～3000
定 价：49.00 元

产品编号：072823-01

前　言

自 2008 年经济危机爆发以来，从美联储开始，世界各国央行都在通过货币政策防止更大经济危机的爆发。大家同时也认识到货币与经济政策只能缓解经济危机的严重程度，争取时间，而不能从根本上走出经济危机。只有新的技术革命的发生，世界才能真正地走出经济危机。

那么下一场技术革命是什么呢？现在有很多推测，如机器人、工业 4.0、虚拟现实（VR）、3D 打印、人工智能等。如果研究一下这些技术的影响范围，就会发现它们都难以担当下一次技术革命的重任。

能把世界拉出经济危机的技术革命是怎样的呢？它应该具有以下特征：①应用领域非常广泛，而不局限在某个行业；②影响的人群非常大，会产生大量新兴的职业；③和日常生活息息相关，具有很大的渗透性。

显然，我们正处于信息技术革命的末期，越来越多的迹象表明，信息技术革命已经完成了它的历史使命：①英特尔公司已经走上了巅峰，摩尔定律即将失效，难以开发出更高性能的 CPU，即使推出更高性能的 CPU，市场需求也不大，市场普遍认为现在的 CPU 已经能满足现有需求；②智能手机经过高速发展后，市场进入饱和期，功能也基本满足需求，大家更换手机的动力减弱，高速发展的苹果公司销量停止了增长；③微软公司推出的 Windows 10 操作系统免费升级，但即使使用多种强制手段，大家升级的愿望也比较弱，甚至认为即使很老的 Windows XP 系统也能满足基本需求。当初人们对信息技术革命的期望，能够做到的已经做到，不能做到的

现在暂时也遥遥无期。

信息技术革命给我们留下什么呢？无论是初期的大型计算机，还是后来的个人计算机、笔记本电脑、智能手机，或者现在很时新的物联网，都会在使用过程中产生大量的数据。这种数据分散在不同的地方，很多数据用完即扔。数据只在产生的过程中发生了作用，历史数据价值并没有被发掘。特别是物联网技术产生以后，收集到越来越多的数据，但很多数据还没有找到用处。

在计算机出现的初期，人们就对利用计算机实现决策支持抱有非常大的期待。关于决策支持系统、专家系统的研究，有一段时间也非常红火，但现在看来这个期望还没有实现。以实现决策支持为目的的商业智能领域虽然积累了许多有价值的理论和技术，但达到的效果还不能满足人们的预期目标。

综上所述，作为信息技术革命成果输出的大量数据应该是下一场革命的输入，是新革命的原料，信息技术革命未完成的使命就是下一次革命的目标。因此，下一场革命无疑应该是数据革命。

数据革命是信息技术革命的延伸，它会对社会产生巨大的影响，它完全是从另一个角度去完成信息技术革命不能完成的目标，还会对信息技术革命的产物（比如计算机网络、云计算、物联网）产生更大的需求，导致现在看来过剩的计算能力又变得不足。

数据革命的影响巨大，会深入到社会经济的方方面面。现在凡是用到计算机的地方，都涉及数据的问题。数据的处理对很多人来说都是一个崭新的领域，有很多新知识需要学习，并因此产生许多新兴职业。

目前已经出现一些与数据革命相关的事情，比如政府开放数据。现在欧美很多国家政府制定了政策，要求政府数据、有政府基金资助的科研数据对外开发，把原来保密的数据变成共享。政府建立起开放数据网站，提供可机读的数据。但是他们的工作局限在数据的供应上，对数据如何利用、如何产生价值还依赖社会的创新，并没有找到通用的方法，也没有找到通用的价值创造机制。

数据革命中的数据不一定是大数据，虽然对海量数据的认知离不开大

数据技术，但大数据只是数据的一个特例。

本书的目的不在于研究数据的获取，因为社会上已有足够的数据，无数运行的软件日夜不停地产生着新数据，无数程序员在编写着程序准备产生更多的数据。本书更多地是放眼于数据时代对数据的存储和应用，以及数据应用会产生哪些改变，这些改变包括政治的、经济的、社会的等；并且探讨了一个通用的数据产生价值的途径——决策支持，其中涉及一个被称为“鹰眼”的核心技术，这个技术的推广应用将对数据使用发挥很大作用。

利用数据进行革命的最终结果应该是：人们通过对数据的分析，解决了在经济生活中遇到的一些难题，反过来推进更多的数据的产生和存储，再进一步推动更多的信息技术产品的生产和销售，吸引更多的人从事与数据相关的工作。



目　录

开 编 故 事

2030 年，中国经济经过深度调整，借助数据革命的春风，走上了一条健康发展的道路。虽然 GDP 的增长率已经大大下降，但由于整个社会发展比较均衡，增长后劲十足。

中国不但拥有完善的基础设施，而且基础设施的收费大大降低，使所有生活、生产必需品在整个生活、生产成本中的占比大大降低，原来家里有空调不敢开的现象已基本不存在了，因为电力成本在收入中已经低到可以忽略不计，所以有很多人即使不在家也整天开着空调。很多人出于环保意识经常在网上呼吁节约能源，但节约能源完全是出于公益心而不是基于成本的考虑。高速公路已基本上实行免费，即使收费也非常廉价，更多的收费是用于不同的时段调节高速公路上的车流，避免高峰时期车流太大导致车速降低、通行效率下降。现在，人力资本已成为所有企业首要考虑的成本。

由于传统的高速公路、高速铁路、电力线路等基础设施建设已非常完善，所以在 2030 年人们的概念中，典型的基础设施已经不是这些东西，而是被称为数据基础设施的数据仓库。通过国家建立的众多的数据基础设施，无论人们在工作还是在生活中都能获取大量既准确又及时、详细的数据，从而大大提高人们的生活和工作效率，一如过去的高速公路提升人们的出行效率。

张开是西部地区一所著名高校的毕业生，读研究生的时候，在导师指

导下开发出一种先进的医疗器械，技术含量很高，市场前景广阔。他在家乡成立了一家公司，并且具备了一定的生产能力，准备到上海去开拓东部市场。

由于准备在上海住一段时间，张开需要租一套房子。张开从上大学开始一直在西部地区，对上海不熟悉，所以他只有在网上寻找相关的信息。他首先打开上海官方的租房数据网站，先研究出租房源和价格的分布。由于他决定租一套两居室的房子，所以从分析不同区域房源数量和房屋租金的分布着手，看到不同租金段的房子在地图上用不同颜色标识出来，颜色密集程度表示房源数量。根据自己的心理价位，张开觉得浦东金桥的房源比较多，房租价位也比较合适。他再分析一下租金历年随时间的变动情况，发现现在是一年中租金的低谷，而下个月由于很多毕业生要求职，每年这个时候租金都会上升，所以他认为这个月应该赶快租好房子。他查询了几套在租房网上的详细信息，比较满意，决定看一下这几套房子，并从中选择一套。

张开之所以对这几套房子的信息真实性没有怀疑，是因为这个平台是政府建立的，统计数据完全是根据房东签约后缴纳税收的数据进行统计，所以数据的准确性毋庸置疑。出租房源的信息虽然是由房东提供通过中介发布的，但中介对信息的真实性负有完全的责任，因为根据国家数据安全法规定，如果发布的信息不真实或者失去时效，一旦租房者为错误或失效数据付出成本，则信息发布者需按租房者为此实际付出成本的十倍进行赔偿。也就是说，如果张开为了租房从西部地区飞到上海而没有租到房，那么信息发布者要按来回机票的十倍价格进行赔偿，或者当天由于信息不真实没有租到房子而住在宾馆，这个宾馆房费的十倍要由信息发布者来承担。如果张开选择入住一家五星级宾馆，即使耽误一天赔偿额也是挺高的。所以信息发布者是非常谨慎的，张开完全不需要担心信息的真实性。

张开在预订飞机票的时候看了一下飞机的航班时间，从他所在城市到上海每天有十个航班，在分析了不同航班的机票价格和误点率等信息之后，他选择了一个很早的航班，因为他想早点将房子定下来就可以不用住宾馆，另外也由于早晨的航班是一天中误点率最低、价格最便宜的。

张开搭乘这趟早班航班到达上海，他在上海很顺利地找到中介，在几套房子中选择了一套满意的租了下来。

在上海的第二天，张开就开始进行市场的拓展工作。因为他的目标客户是医院，所以他要找到上海所有医院的资料。在2030年，由于数据基础设施的发达，以前的人脉已经失去作用，如果像以前要依靠人脉的话，像张开这种在上海举目无亲的人是无法在上海开拓市场的。而现在的政府数据平台上，公开了上海所有医院的信息和数据。张开先分析了一下所有相同功能医药器材的采购、使用统计和增长情况，发现使用量庞大而且增长迅速，所以他对做好这个市场很有信心。然后他又在这些数据中，找到使用量排名前三的三家医院，准备从这三家医院着手。由于他的生产能力有限，他准备先打开市场，前期先占据10%的市场份额，回去再慢慢扩大生产规模。在数据平台上，张开除了看到使用量以外，还看到每家医院采购的平均价格。他认为他的产品除功能有创新外，价格还是有竞争力的。他的价格比现在的采购价格大约便宜10%。

张开从网上顺利找到这三家医院采购部门的联系人和联系电话。他打电话过去，和第一家医院顺利地约好见面时间。第二天按照约定的时间，他拿着自己的样品到达那家医院，拜访了这家医院的采购负责人。这家医院的采购负责人是一名非常内行的专家，在看到样品之后，对产品的功能和质量非常认可，当即决定先采购小批量产品在医院进行试用，满意后再扩大采购规模。

首战告捷，张开非常高兴。回去之后他和第二家医院也约好时间，前去拜访。这家医院的采购负责人并非专业人员，对产品的质量难以把关，当他看了样品之后对产品质量产生怀疑。但他为人很好，建议张开到政府的检测中心检测一下产品质量，并明确表态，若是质量没有问题他就会采购小批量试用。于是张开根据他的推荐，找到一家检测机构，将自己的产品样品送过去检测。检测机构在三天之内给他产品的检测结果。张开将检测结果拿到医院之后，采购负责人也同意试用。张开承诺交付产品的质量和检测结果是一样的，如果他的产品和检测样品有差异的话，根据质量法，他将会承担巨额的罚款，他的公司也可能由于这项罚款而倒闭。所以第二

家医院看到他能够生产出这个样品，并不担心批量产品质量的稳定性。

第三家医院的采购负责人开始拒绝和张开见面，也不说理由。张开把自己产品的功能优势、质量检测结果，根据该医院使用量统计数据计算的成本节约数据告诉这名负责人，并表示将向政府公平竞争管理部门投诉，该负责人才答应过几天见面。张开后来才知道该负责人家里遇到事，心情不好。

张开正是由于在开拓市场方面所费功夫甚少，所以他把精力主要放在产品功能的完善、生产质量的保证和产量提升上，根本不需要在销售方面下太多功夫，而且他也知道随着其产品的供货扩大，其他供应商也会在产品的性能和价格上和他竞争，也就是说，他现在的优势只能维持一年左右，所以他还必须花大量的功夫在新产品的研发上。

张开的东部市场开拓异常顺利，他给一直向他泼冷水的舅舅打电话报喜。舅舅对他这么快就能打开市场觉得不可思议："数据基础设施建设效果真的这么神奇？"

张开对未来充满信心，计划在三年内让公司上市，并走向国际市场。



第 1 章

迎接数据革命

信息技术与经济社会的交汇融合引发了数据迅猛增长，数据已成为国家基础性战略资源，大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。

国务院《促进大数据发展行动纲要》

1.1 信息技术革命

数据成为重塑国家竞争优势的新机遇。在全球信息化快速发展的大背景下，大数据已成为国家重要的基础性战略资源，正引领新一轮科技创新。充分利用我国的数据规模优势，实现数据规模、质量和应用水平同步提升，发掘和释放数据资源的潜在价值，有利于更好发挥数据资源的战略作用，增强网络空间数据主权保护能力，维护国家安全，有效提升国家竞争力。

国务院《促进大数据发展行动纲要》

1.1.1 未完成的第三次工业革命

18 世纪中叶以来，人类历史上先后发生了三次工业革命。目前，大家一致认为第三次工业革命已经结束，并期待第四次工业革命的到来。

有人把第四次工业革命，定义为以互联网产业化、工业智能化、工业一体化为代表，以人工智能、清洁能源、无人控制技术、量子信息技术、虚拟现实为主的全新技术革命。显然，一次工业革命不可能是一些新技术的杂烩，而且这些新技术还依赖数据技术的发展，比如工业智能化，在数据认知还没有完成时，智能决策和执行无从谈起。

从第一次工业革命的蒸汽机和第二次工业革命的电力设备来看，蒸汽机的发明、制造、安装和维护虽然是一个巨大的产业，但产生的社会影响远不如其带来的规模化生产（比如钢铁厂和纺织厂的生产以及铁路的铺设）来得深远。同样，从电力的发明来看，电力本身有发电机以及发电机的生产、制造和服务，但它的影响不如后来电力输送线的铺设和大量电器的使用以及生产中动力从蒸汽机转变为电力带来的影响大。

综上所述，每一次工业革命都由两个或多个部分组成：第一个部分是作为引起革命标志的技术的发展；第二个部分是由这个革命的技术导致的社会更深层次的影响。

同样，我们来审视一下第三次工业革命。第三次工业革命是以计算机的发明使用为基础，计算机本身已经产生了一个非常大的产业，对人类社会也产生了巨大影响，但这不是第三次工业革命的最终结果。计算机产生的大量数据以及由于我们对数据的应用而产生的变革才是第三次工业革命更大的成果。

数据革命也不是人们传说中的第四次工业革命，只是第三次工业革命的下半场。第三次工业革命分为两部分，上半场是信息技术革命，下半场是数据革命。原因有两个：一是这两场革命的联系如此密切，难以分开；二是数据革命实现的是信息技术革命未完成的理想。

1.1.2 从智力替代到辅助决策、自主决策

蒸汽机发明的重要意义，在于人类首次从只能依靠人力或者畜力作为动力，变为可以以机器作为动力，从而对人类的生产经营活动产生了巨大的影响。原来只能小规模生产的产品因为机器的发明能够进行大规模的生产，火车头或者轮船可以通过蒸汽机来驱动把人或货物运送到很远的地方。

计算机的发明，同样拥有一个相似的重要意义，人类首次以机器来代替人类的智力活动。蒸汽机以机器代替人类的体力活动，计算机以机器代替人类的脑力劳动。

但是，仔细分析从计算机发明以来人们所取得的成就，不难发现，计算机的发展还没有完全达到预期的目标。

机器代替人类智力的活动有三个阶段。

第一个阶段是智力替代，即原来是人可以实现的智力活动，用计算机来替代。比如，原来必须用心算或者笔算进行的计算，用计算机可以自动进行；原来必须依靠个别智力超群、经验丰富的人才能完成的工作，可由计算机来完成。第三次工业革命基本上完美地实现了这个阶段的目标。

第二个阶段是辅助决策，可以简单理解为人类的决策活动由计算机来提供辅助。虽然人类没有计算机的辅助也可以自主决策，但由于精力、时间等诸方面限制，决策的质量通常并不理想。比如，指挥员在不知道敌情的情况下，也可以发起一场战役，但常常会遭遇惨败。决胜的关键在于能否搜集到足够的情报。企业管理中能否做出正确的决策，主要看是否准确掌握市场和自身的综合信息。就是说，辅助决策能提升人类智力活动的水平，就这点来说，现在还没有达到要求。

第三个阶段是自主决策，即机器可以在没有人工干预的情况下通过自主学习进行决策。现在以深度学习为核心的人工智能已经能做到自主决策，但毕竟这个发展才刚刚开始，还有很长的路要走。换句话说，这方面已经有了起步，但还远远没有成熟。

后两个阶段的需求并不是现在才提出来。在计算机刚发明的时候，人们已经提出这个需求，包括基于辅助决策的DSS（决策支持系统）的研究、AI（人工智能）的研究，在几十年前都已经开展起来，并且取得一些成就，后来由于技术的局限而停滞了很长时间。以人工智能为例，在深度学习的算法得到突破以后，才有新的发展。同样，辅助决策也在技术上陷入了停滞，只有引进新的技术、新的思路，才会得到发展。

综上所述，第三次技术革命是一次信息技术革命，可以分成三个阶段：智力替代、辅助决策、自主决策。也可以分成两个时代：信息技术革命时代和数据革命时代。现在完成了信息技术的革命，只完成了三个目标中的智力替代的工作，后面两个工作需要在数据革命中完善。

1.1.3 三次工业革命的比较

前文介绍了三次工业革命，接下来对三次工业革命的过程进行认真的分析，以此为借鉴来评估第三次工业革命以及未来的发展。

第一次工业革命虽然以蒸汽机的发明为标志，但实际上它真正对社会产生的巨大影响在于后来汽船的发明和铁路网络的建设，后者从根本上改变了人们交通的手段，使人和货物的来往更为方便廉价。比如：汽船的发

明使欧洲到美国的大西洋航行更为快捷，从而使得大量的移民可以抵达美国。火车的发明，特别是美国太平洋铁路的建设使美国东西海岸得以相连，大大加快了美国西部的开发。

第二次工业革命虽然是以电力的发明和使用为标志，但其巨大影响和电灯的发明与电力网络的建设密不可分，就是说只有在电力网络建设比较完备，电力能被很多地方的人所使用的时候，第二次工业革命才真正地发挥了作用。

目前，中国有三个非常大的垄断企业，一个是铁路总公司，拥有中国整个的铁路网络，还有两个就是国家电网和南方电网，垄断了中国的电力线路。它们分别是两次工业革命的成果，即使从现在的角度来看这两个网络也是一个国家经济发展的基础，其重要性有目共睹。相对而言，火车机车制造企业和发电企业的重要性就差很多。

第三次工业革命发展到现在，虽然我们有了计算机和互联网，但它和铁路运输及电力传输的差异是明显的。现在在互联网上传输的都是用 HTML 标准标记的语言制作的网页，相对于我们拥有的数据，可以在网络上传输并且识别的数据显然数量还很少。

对比铁路和电力网，铁路运输的是人和货物，它只负责将人和货物从一地转移到另外一地即可，这是一个通用的运输工具，不管货物和包装是什么都能送达。同样，电力输送的电输送到任何一个地方都能够被任何以电力为能源的设备所使用，也是一种标准化的产品。

互联网和铁路、电力网有比较大的差距，虽然互联网的网络已经铺就，但传输的数据没有标准化。传输的数据从一地到另外一地，并不能被人们方便地采用，而必须通过专业的协议和手段才能看到。这些信息的格式比较多，包括文字、图片、音频和视频等。任何一个人打开一个数据包，并不能保证他能读懂数据，这是因为有很多不同的数据格式。

因此，三次工业革命结果的差异就在于：第一、第二次工业革命是先有标准化的产品再有网络的建设，第三次工业革命是先有网络的建设后再有标准化的产品。那么，要完成第三次工业革命还需要什么呢？就是还缺乏一次数据革命，实现把数据当成标准的产品来传输。

大家知道，在一个新产业兴起的时候都是百花齐放，有很多的标准一起出现，有一些混乱。但当一个产业成熟以后，技术指标总是归于一个标准。所以，目前在网络传输的信息混乱正是信息革命初期的一个标志。在数据革命完成后，类似以 HTML 语言为标准的网页会扩充到数据上，一个人用浏览器就可以阅读不同的数据源提供的数据。

1.1.4　数据是信息革命的主要遗产

自第一台计算机发明以来，信息技术革命取得了巨大的成就，使人类第一次能用机器代替人脑，就像第一次工业革命，让机器第一次代替人力和畜力一样。

计算机可代替人脑或者一些辅助手段，比如计算尺、算盘等工具来计算，实现了用机器代替人脑计算，从而使复杂和大规模的计算成为可能。人类只需要把计算的过程编制成程序，无须每次重复相同过程，就可以由计算机得到计算结果。

在我们所处的时代，我们的很多行为都深深地打上了信息时代的烙印。没有计算机、互联网和电子邮件这些工具，经济全球化无法实现。很多新式武器及航天器，无不建立在信息技术发展的前提之下。智能手机的普及不但使人手一台计算机得以实现，而且使信息技术的受益者从拥有专业技能的人员走向普通大众。

回顾信息技术几十年的发展，计算机本身发生了哪些变化呢？

计算机第一个变化是从无到有；第二个变化是从大到小；第三个变化是从单机走向联网。

计算机从占几间屋子的大型计算机，到放在一个房间的小型计算机，再到桌面上的 PC，又到人手一部的智能手机，最后变成米粒一样的物联网的智能传感器，其体积越来越小。

计算机从单机走向联网的发展：从原来的一个单位一台计算机到一个部门一台计算机，再到人手一台计算机。一辆汽车上有几十台计算机，未来随着智能家居的发展，一个人家里将有好几十台计算机。

计算机的联网使计算机中的信息可以共享，每台计算机都不是一个独立的存在。自己的数据可以被别人读取，同样地也可以读取别人的数据。如果没有互联网，计算机的价值就会大打折扣，因为有些计算机的功能就是读取网上的信息，而自己根本不产生任何信息。

计算机的发展还使原来人类认为完全不同类型的信息全部变成数据。人类大脑可以认知的信息包括视频、音频、触觉、嗅觉和味觉，视频信息和声音信息都实现了数字化。原来用各种各样模式存储或者传输的信息全部变成数据。因此，现在社会上产生最多的就是数据，保存最多的也是数据，以后任何一个人的生活中都离不开数据，所以说数据是信息技术革命留下来的最大资源，以后还会不断地增加。如果我们不能很好地处理数据，不但不能为人类服务，可能还会产生一定的灾难。所以，数据革命是继信息技术革命之后的又一次机遇，更是一次挑战。

1.2 为什么是数据革命

这次发生的不是第一次数据革命。

在数据的发展历程中有过两次数据革命。第一次数据革命是近代科学诞生之时，实现了数据与科学研究的融合，数据在科学研究中的基础地位得到确立。对研究过程和结果赋予精确化的诉求，是近代科学的基本特征之一。在以数据为依据的研究范式中，数据的可靠性和准确性代表了研究的精确性，人们甚至将以数据为依据的实证研究作为判断“科学”与“伪科学”的标准。[1]

第一次数据革命解决的是从无数到有数问题；第二次数据革命解决的是从小数到大数的问题。

为什么说是数据革命，而不是数据改革或者是数据技术呢？

首先，革命有它的界定条件，它的影响面要足够广。比如工业 4.0，它只局限于工业，对医疗、教育并无大的影响，故而不能称之为革命。其次，

革命的深度以及对时间的要求：革命不可能在短时间内完成，它会持续很长时间，甚至产生的影响力会持续几十年乃至几代人。革命需要众多人参与，并非一个公司开发出一个产品就能够作为一场革命。革命需要非常多的社会资源，从政府到个人都要积极参与才能产生效果。革命要对经济产生巨大带动作用，能够引导整个社会资源的配置改革方向并带来投资。

除此之外，革命不可能凭空发生，它需要在原有技术的发展基础上延续下来。同时，革命需要有非常明显的标志，它不是一个渐进式的事物，而是一个有独立特征和广泛影响的事物。

如今第四次工业革命随着科学技术的发展在不断孕育中，工业 4.0 是否就是第四次工业革命呢？我们都知道信息技术革命（第三次工业革命）的主要特征是计算机的发明和应用，以及互联网的发明和应用，其产生的影响大家有目共睹。但是，信息技术革命到现在已经逐渐进入尾声，特征体现在计算机的应用率越来越高，特别是智能手机的发展和普及，由于其功能和计算机的功能越来越接近，现在已经达到人手一台计算机的水平。

此外，因特尔公司的 CPU 开发使计算能力已经远远超过了现有的需求，所谓的摩尔定律物理上已经达到极限，智能手机的销量也已经从顶峰逐渐呈下降趋势，所有的一切都证明信息技术革命已经结束。

但是，我们有些问题在信息技术革命中并没有得到解决。在计算机刚刚发明的信息技术革命早期，就有人提出了 DSS 的概念，可是到现在信息技术革命快要结束了，人们并没有达到在信息技术革命初期的预想。现在的技术还有很多的局限，用现在的技术和思路并不能解决这些问题。BI（商业智能）技术已经开发并推广多年，却迟迟不能得到普及。

信息技术革命给世界带来巨大进步，也留下很多问题，主要集中在数据的共享和利用上，而数据革命将解决其中的大部分问题。

1.3 社会需要数据革命

1.3.1 发展需要资源配置均衡

资源配置均衡既是社会公平正义的要求，更是经济快速发展的前提。资源指的是资金、土地、原料、能源、教育、医疗等。一个国家无论贫穷还是富裕，只要资源在不同地区、不同人之间分布是相近的，就处于均衡状态。与均衡相反的是贫富不均，甚至贫富悬殊。

经济的发展都有商业周期，都是从均衡走向不均衡。不均衡的产生是由于在经济发展中，不同行业有不同的发展规律和发展周期，有的行业遇到技术突破或市场机遇，会得到迅速发展，吸收大量社会资源，有些人由于天赋、家庭、教育、个人经历、从事行业等原因富裕起来，比其他人占据更多资源。任何一个区域或时代的经济发展，都会经历从不足到过剩，最后到泡沫的过程，在泡沫阶段不合理地占用过多资源，需要调整。泡沫的破裂使这些资源被释放出，可以被其他新兴行业吸收，走向新的均衡。因为无法预知下一个经济发展的机遇在哪里，处于均衡状态的资源最容易被新的机遇所吸引，就像在平原上的水可以向任意方向流动，而位于山地的水流动就会受到群山的阻碍。如果泡沫不破裂，处于这个行业的资源就无所适从，不离开会觉得前途渺茫，离开又会觉得这么多年积累丢弃太可惜。

中国三十多年的经济快速增长，创造了人类社会的奇迹。对其中的原因，有不少专家做了研究，甚至提出了“中国模式”。改革开放初期的资源配置相对均衡状态，是中国经济高速发展的主要原因。

中国在改革开放之初，虽然经济落后，但由于新中国成立后的土地革命、公私合营等，使原先的地主、资本家不再占有过多资源，整个社会处于均衡状态。在政策开放，全国集中精力发展经济后，整个社会发展动能十足，中国经济得以腾飞。

对中国而言，政策至关重要，邓小平功不可没，但从世界范围来看，

仅有政策是不够的。很多国家几十年来一直致力于发展经济，但发展一直很缓慢，原因就在于国家内部存在严重的贫富悬殊，未处于均衡状态。在解释拉美和北美巨大不同时，专家认为在北美可供剥削的土著比较少，资源配置比较均衡，而南美一开始就建立在少数人剥削广大土著之上，资源配置不平等问题迄今都没有解决。

中等收入陷阱产生的原因是在经济得到一定发展后，资源配置均衡被打破，既得利益行业、企业和个人占据过多资源，即使在经济发展停滞、资源过剩的情况下，也不愿意将这些资源释放出来，政治和经济体制上没有建立像美国一样的调整机制，最终导致经济失去活力。目前中国也面临同样风险。

“二战”后的德国、日本经济的快速发展，也是得益于被战争摧毁后的平衡状态。特别在日本被盟军占领期间，盟军司令部强制解散财阀，并指令日本政府制定法律防止垄断资本复活，使日本社会资源配置得以均衡。

美国经济相对欧洲、日本而言，具有很大弹性，比如目前美国就率先从经济危机中复苏，这与美国社会的再平衡能力有关：美国公司遇到经济低迷就裁员、破产，经济回升就扩大规模。

日本在 20 世纪 90 年代，虽然刺破了房地产泡沫，但维持了企业泡沫，很多效益不好的企业通过银行输血活了下来，没有完成再平衡，所以社会缺乏活力，新企业少，年轻人就业困难，导致现在的经济困局。

在市场经济下，市场的主体基于个体利益的考虑，会让资源配置向不均衡方向发展，比如房地产热的时候，资源会向房地产业聚集，除非泡沫破裂。政府的作用应该致力于均衡，所有行为应该有利于资源在全社会的均衡配置，而不能助长不均衡的倾向。

数据革命的目的是使政府和全社会能够掌握资源分配的状况，防止资源浪费，在泡沫产生的时候及时预警，或在泡沫破裂的时候及时调整资源配置，给政府的调节指明方向，对调节的结果及时予以反馈。

1.3.2　数据促进社会平等

资源配置的均衡建立在平等的基础上，平等分为机会的平等和结果的

平等。结果的平等是一种不公平的平等，如果实现了，实际上是平均主义，吃大锅饭。机会的平等，是资本主义推崇的一种平等。如何实现机会的平等呢？目前被认可的主要做法是保证教育的平等，就是保证一个人不管出生背景如何，只要自己有天赋并努力，都可以受到良好的教育。实际上，即使接受同样教育的同学，家庭背景对一个人的成功还是有着重大影响的，这种影响的原因之一是由于背景的差异，每个人获取的信息不对称，从而机会也不对称。

随着信息技术的发展，越来越多的信息来自数据，所以数据的平等就是信息的平等。数据的平等，应该是机会平等的重要组成部分，是和教育平等同样重要的平等。数据的公平，就是社会上每个人都可以公平获取个人发展所需要的数据，比如专业的报考人数、毕业人数、工作薪酬，招聘的企业和岗位数、工资水平、所在行业和区域等，从而一个人无论是就业还是创业都有公平的起点。

1.3.3 不均衡导致中国古代王朝更迭

数据是衡量社会财富是否均衡的有力工具。

中国有五千年的文明史。从秦始皇开始的专制历史就是一部王朝不断更迭的历史。很多学者都研究过中国古代王朝更迭的原因，一般认为是由于最高统治者采用世袭制，后代皇帝养尊处优、治理能力下降而引起的。

中国古代每个新王朝基本都是建立在战争基础上。累年的战争导致生产力的破坏，原来占据较多社会资源的人由于死亡、迁徙、社会地位的变化等原因失去原有的优势，社会资源的分布重新变得较为均衡。

但随着时间的推移，新的强势群体逐步掌握了越来越多的社会资源，积累了越来越多的财富，社会资源开始向这少部分人集中，于是社会开始逐渐进入新的不均衡状态。这种不均衡开始并不太严重，人们还可以接受。但经过几百年，这种不均衡逐步发展到忍无可忍、民不聊生的状况。这时，农民起义就会爆发，开始了新的改朝换代进程，直到新王朝建立，进入新的均衡状态，再开始新一轮的循环。

从明朝的封藩制度，可以看出随着时间的推移，不平衡是如何逐步发展到触目惊心的程度的。

明朝分封诸皇子为亲王，并规定了一套严格的封藩制度。到了嘉靖初年，明朝的宗室总数就比明初膨胀了上千倍。万历年间，总数增长到三十多万个，明末天启年间，更有六十多万个。天下供应京城的粮食，每年400万石，但各王府消耗国家的粮食，每年却有800万石。具体到地方上，军事重镇山西省，每年存留粮食190万石，但当地王府消耗粮食，却有300多万石。河南省存粮94万石，当地藩王消耗粮食，却有190多万石。也就是说，全国的税粮加起来，也填不满藩王的嘴。[2]

一个社会，如果拥有足够准确的数据，有利于及时发现社会不均衡状态，当政者也因此可以及时调整。试想当年如果明朝历代皇帝可以对宗室总数、消耗粮食数量、消耗粮食与全国产量的占比进行分析和预测，就可以发现问题的严重性，及时进行调整。实际上，由于宗室数量增加是个缓慢的过程，明朝皇帝有几十年到上百年、几位皇帝的任期时间去做调整这个事。

1.3.4 熵增原理

在物理学中有一个与能量守恒定律具有同等地位的基本定律——熵增原理。在一个相对封闭的体系中，表达混乱程度的衡量指标“熵”在没有外力作用下是一直增加的，除非有外力的作用“熵”才会减少。就是说，一个封闭的系统都是自发从一个相对平衡的系统转向一个较为混乱的系统，这种变化是自发行为，无须人为干预，而只有需要回到平衡状态时才需要外力干预。

将此原理运用到经济上，得出的结论是经济发展的自然发展方向是从平衡到不平衡。

如果一个社会的财富分配比较平均，社会各种资源的分配比较合理，那么这就处于一个熵值较小的状态。在没有任何外力作用，比如政府调控或天灾人祸等因素的作用下，它会自然地向熵增的方向发展，进入一个贫富不断分化、资源不断错配的状态。这个状态不需要人为干涉，它通过市

场机制就能自然发展。

在熵值增加到一定程度时，贫富悬殊、资源错配会严重阻碍经济的发展，经济发展会停滞不前，直到经济危机爆发、战争爆发或改朝换代。

为避免经济发展的停滞，需要进行有利于“熵减”的干预，让经济向平衡状态发展。只需在熵增过程过于剧烈或者过于突出的地方进行干预，从而实现“熵减”方向的调整，无须随时干预。顺着“熵增”的方向干涉不仅无效，反而会加剧它的不平衡。

数据革命的目标，就是计算出经济的熵值。通过对全社会经济数据的分析和研究得到熵值，然后从全局或者某个行业观察熵值的变化，用适当的方式进行干预，减小熵值或者减缓熵值的增长速度。通过观察干预后熵值的变化，得到干预措施的反馈，知道干预的效果，对干预方法进行微调。当然，如何计算衡量经济的熵值将是一个巨大的挑战。

1.3.5 中国国内市场的完善

有专家对历史上大国在对外开拓市场上不同的做法进行了分析，提出了一个适合中国的战略：以优先开发国内市场来带动“一带一路”战略的成功。

文章的核心是把历史上大国开拓市场的模式分为四类：第一类是抢占现存大国的市场开拓模式，如德国，这是一种危险的模式；第二类是以日本为例的依赖霸权国家的市场开拓模式，这是一种比较脆弱的模式；第三类是英国的靠边缘国家的市场开拓模式，这是一种比较摇摆的模式；第四类是以美国为例的基于国内开发的市场开拓模式，这是相对比较稳固的一种模式。[3]

从中国的经济来看，前三十年的改革开放主要靠出口，所以虽然劳动力和生产工厂都在国内，但实际上是完全按照国际市场的规则在生产、贸易。现在由于劳动力成本上升，出口市场萎缩，不少企业把市场转向国内。当转向国内市场后，发现国内市场贸易规则和国际市场完全不一样，很多方面不成熟，这些企业面对的是一个全新的市场，需要遵守全新的规则。

美国的一些成功企业可以把自己的商业模式在世界范围内复制，但到

了中国就不行。通常把这种失败归咎于美国企业缺乏灵活性，而实际上是因为遇到完全不同的市场规则。

中国国内市场规则的特殊固然可以保护国内企业，但同样也会阻止中国企业走向国际市场。

最理想的方法是按照国际市场的规则去重塑国内市场。然而，因为存在着很多的制度和现状的瓶颈，短期内难以实现。但如果按照德国模式跟现存大国抢占市场的话，则可能会发生冲突。

比较切合实际的思路，是按照中国改革开放的成功经验，存量不变做增量，不急于在固化的现有体制上动刀，而是建立新的体制。现在国内市场有很多壁垒，如政府管控的壁垒、运输成本的壁垒以及资金流动的壁垒等，这些壁垒都阻碍了国内市场一体化的形成。但是基于数据的信息共享壁垒还没有形成，在这种情况下，要大力建立数据的共享机制。

数据共享的主要特征是，建立公益性的、共享的数据交换机制。在数据壁垒还没有形成之前，必须抓紧时间建立。等有相关的利益集团形成以后，再建立就很困难，改革会遇到很大的阻碍。

现在，有一些公司已经在试图建立数据共享壁垒了。例如，阿里健康在医药市场对医疗数据建立监控码的平台时，就在试图建立这样一个壁垒。这个平台虽然名义上是国家食品药品监督管理局拥有，但是所有的经营权都归公司，政府在技术上缺少话语权，数据被阿里健康垄断，在数据共享上以安全为由设置障碍。此案例，表明政府对数据共享缺乏长远规划，没有数据共享平台的机制设计。

1.3.6 新的就业机会

数据革命能够带来新的就业机会。

现在的经济危机需要一场技术革命才能带来复苏。这种技术革命的最终结果可带来大量的就业机会。这种就业机会主要集中在几个方面。

（1）推动信息技术进一步发展。现在的数据虽然可以采集，但人们不知道怎么用，所以让人觉得数据有多余。数据的采集更多地由技术驱动，

比如物联网。物联网的概念兴起以后大家都在搞物联网，很多人在研发物联网设备，推广物联网的应用，但应用以后产生大量数据怎么办，结果发现用处不大，或有些很小的用处。到了数据时代，等发现了数据的应用方法后，就会发现现在的数据远远不够用。因此，对信息技术、信息设备及其相关软件的开发会产生大量的需求，现在的信息技术的应用和设备的推广出现新的、更大的发展空间。

（2）有许多信息需要转换成数据。比如耶鲁大学的陈志武教授在国内合作搞的量化历史，就是把历史记录中的一些事件量化成数据。因此，信息时代以前的很多资料都可以做量化，目前很多信息没有数字化。在很多研究领域，研究的方式会产生很大的变化，对信息技术产生很多新的需求。医学从中医的辩证，到西医的手术，再到基因组的数据，基因组技术本质上可以说是数据技术，基因组的工作实际上就是把组成人体的基本信息进行数字化。按照这样的思路，有很多行业都会出现这方面的需求。

（3）对数据的转换和保存。现在数据都分散在不同的地方，以不同的格式存储，以后要像挖掘文物一样把它们挖掘出来，让它们发挥作用，这样会产生大量的数据处理工作。

（4）数据分析。对数据怎么解读，怎么预测，需要大量掌握数据分析技术的人员。

所以数据革命产生许多新职业，需要许多新的从业人员。

1.3.7 建立社会经济运行的反馈机制

反馈是物理学中非常重要的概念，若没有反馈很多的物理系统根本无法正常运行。

人类走路也离不开反馈。如果我们想沿着一条直线走，并且把眼睛蒙上的话，就会发现人实际在转圈子。我们在走路时，实际上需要眼睛不断反馈信息，不断修正步伐，最后才能走出一条希望的直线。

人类的工作只有得到及时、有效的反馈，才能做得更快更好。奖励也是反馈的一种。拿破仑说：“只要有足够的勋章，我可以征服全人类”。

在基础设施建设等领域，可以很方便、直观地获得反馈信息，比如建设一座高楼，每多建成一层，目视即可了解进度情况，但在更多情况下，只有通过数据才能得到反馈信息。

在社会科学领域，就缺乏很好的反馈机制。比如，现在虽然有统计局的数据对宏观经济进行一些反馈，但实际上数据不太准确，而且比较滞后。

1.3.8 权威的信息交换平台

许多新兴产业的发展需要一个权威的供需信息交换平台作为支持。目前，很多的数据发布在不同的平台上，缺乏权威性，需求者需要去不同的平台查询，对数据的可靠性也有怀疑，这一切影响了数据经济的发展。我们需要一个统一的数据操作平台，并且保证数据的权威性，目前可行的方法就是这个平台必须是公益性的，必须是由政府或者一个公益组织提供的，它不提供具体数据的变现，只提供原始数据，并且保证数据的合法性。数据的合法性需要法制的保护，比如在数据时代可能需要对数据发布的准确性立法，如果你发布了虚假信息或者信息过期没有及时清理，需要承担一定的法律责任，或者对这些信息导致的后果要承担赔偿责任。举个例子来说，有一个人发布了一条虚假信息，哪个地方有房子闲置出租，而另一个人知道这个信息后打车或坐飞机去现场看房，结果到现场一看这个房子已经出租。这种情况下求租者可以起诉信息发布者要求赔偿他的交通费用和误工费用。由于这个事情涉及的总额比较小，这里可以设一个比较高的惩罚金，比如按实际发生费用的 100 倍赔偿。另外，对使用数据平台的商业公司，在业务完成以后，有义务随时协助发布者清理这个数据。如果没有更新、清理数据，也要承担相应的法律责任。在公共数据平台上，数据发布者对数据的准确性负有法律责任，但是，一旦有商业公司接过这个数据提供服务以后，该公司有义务对数据的状态进行及时的更新，否则应该承担相关的法律责任。

1.3.9 分享经济模式的扩张

从 Uber 的商业模式及它所包含的意义，可以看到发达西方国家虽然还不具备数据时代的基础设施，但有公司已经起步。

Uber 是美国一个公司推出的新型打车业务。一般的出租车有专门的公司、专门的车辆和专门的司机来提供出租服务，有政府专门的定价，它是一种被政府法规规范的职业。但是 Uber 提供了一种兼职的行为，只要你有车，有空余的时间，其他人需要打车的时候可以通过 Uber 公司的平台叫到你。

Uber 的成功实际上一个是信息技术的成功，因为它的前提是打车的人和 Uber 司机都拥有智能手机，智能手机提供了人与人相连的网络功能，而且智能手机还有一个重要的定位功能，这样就能够非常方便地让 Uber 司机了解顾客准确的地理位置。试想一下，如果没有智能手机的普及和定位功能描述，那么一个在大路上的行人必须使用计算机，而且必须精确描述地理位置，而 Uber 司机必须在自己车上配备电脑设备并能上网才能进行业务处理，这样的话，双方成本肯定都很高。

现在由于智能手机的发展，对计算机的定位、上网功能都不需要额外的花费，利用现有的功能就行了。也就是说，一个 Uber 司机，他根本不需要做任何的投资，加入 Uber 网络就可以了。当然 Uber 的成功还依赖他跟美国政府监管部门的不断斗争。因为这种模式违反现有的法律法规，它是打了很多的官司以后才得以成功运营。中国的“滴滴出行”模仿了 Uber 的这种模式。

从表面来看，Uber 通过信息技术实现了对闲置资源的合理配置，把原来闲置的车辆和闲置的人的时间利用起来，为社会增加了出租车服务，应该说是一种促进资源优化配置的先进技术，因此 Uber 模式得到非常高的评价，被认为是该领域的一场革命，甚至在其他行业遇到类似挑战时被称为“Uber 时刻”。

通过深入分析，发现 Uber 模式还有更深层的意义：它实际上是数据时代的先驱。

从数据角度看，Uber 实际上是收集了车辆和打车需求数据，然后进行了匹配，最后产生了业务。因此核心是一种数据业务，也就是说，你只要把需求数据和供应数据通过计算机的匹配，就可以产生相关的收入。

Uber 的数据使用的方法实际还是一种传统的数据库检索方法，跟真正数据时代的主流技术方法还是不一样的。所以 Uber 只是一个数据时代的开始。

Uber 模式有什么局限呢？虽然 Uber 实现了数据服务的功能，但为了实现这个功能，它跨越了两大门槛。第一个门槛是法律法规，它跟当前社会的机制做了很多的斗争。如果不跟政府打官司，它根本开展不了现在的业务。它在美国各地的业务都是一个州、一个市打官司争取下来的。这反映了现在的政治体制并没有为通过数据配置资源提供足够的支持。第二个门槛是 Uber 利用了大量的资金，Uber 成功的非常重要的一点就是以不断提高估值来融资。Uber 的商业模式并不是它发明的，而是另外一家公司发明的，这家公司由于没有像 Uber 这样疯狂的融资已经倒闭。也就是说，需要大量的资金支持这种数据的应用。

通过 Uber 案例可以看出，人类要进入数据时代非常困难。怎样降低数据时代的门槛，政府应该做哪些工作，让类似 Uber 的模式四处开花呢？

Uber 现在开展业务的数据是自己花巨资宣传、推广采集的。如果这些数据能由政府资助的公益平台拥有和发布，任何公司免费使用，那么类似 Uber 的公司就可以低门槛进入。该方案把 Uber 的模式一分为二，一部分就是数据的采集服务作为公用事业；另一部分对这些数据的应用作为一个私营企业的核心竞争力。

Uber 在打车领域是成功了，但按照现在的模式，在许多的其他领域很难出现 Uber 这样的公司。因为作为一个公司它要专注，不可能轻易进入其他领域；第二它要找最容易实现的商业模式，打车这个需求促进了 Uber 的成功。如此推广开来，在其他需要 Uber 这种模式的领域，不大容易出现 Uber 这样的公司，要在其他领域推广很难。但如果把 Uber 的模式分成两部分，在其他行业相对就比较容易复制了。

实际上在美国不止出现 Uber 这一种模式，还有两个公司也是很成功的，

成为了所谓的“独角兽”公司。一个叫 AirBnB，另一个叫 WeWork，他们的模式实质和 Uber 是一样的。

AirBnB 让人们把家里空余的房子在它的系统上进行登记出租，然后要租房就可以住到人家家里去。这个房子可以是整栋房也可以是一个房间，就是把现在专业的宾馆提供的住宿服务延伸到家庭，就像 Uber 把专业出租公司延伸到普通的汽车拥有者身上一样。AirBnB 实质上把房屋资源和时间资源拿出来共享。

WeWork 的模式要稍微复杂点。它是把空余的办公室、办公楼租下来，经过改造，然后分租给不同的需求者。原来租办公楼只能租固定面积的一大间或者一层，但如果人员在迅速扩展中，比如原来是 10 个人，现在发展到 100 个人，就需要不断地搬办公楼。如果人员很少，只要很小的办公室，甚至只需要一张桌子，可能就没有地方可以提供。而 WeWork 是把办公楼租下来以后，分割成小房间甚至单个办公桌，让你根据需求去租，一个月、一年都可以，当人员增长时，只要扩大租赁就可以了，比较弹性地满足了租赁者的需求。空余的办公楼业主都希望整层、大间出租，出租不掉可能就空置在那里，而 WeWrok 通过改造以后充分利用了这些资源，满足了这些小型的或者快速发展的公司的需要。

Uber、AirBnB 和 WeWork 的核心理念是相似的，主要通过数据服务，实现资源的充分利用。当然 WeWork 要复杂点，需要把房子租下了并且进行装修，Uber 和 AirBnB 相对简单一点。

1.4 从海关数据看数据价值

海关数据是世界各国海关的货物进出口交易数据，主要来自关单、提单、商检等真实的单证记录。海关数据信息是国家掌控国际贸易变化、制定贸易政策最直接的依据，同时也是企业市场调研、国际市场开发的重要资源。

海关数据公开是合法的。根据各国的政府法令、贸易政策的不同，每个国家对海关数据开放的程度不一。其中开放程度最高的是美国，早在19世纪70年代美国就通过政府立法的形式公开海关数据，后来授权给专业公司进行商业化的运作。中国海关也有相关规定，海关资料可以公开。

海关数据的详细程度超过大家的想象，一个公司如果出口一批货物，就会暴露许多商业秘密。

以下是一个关单的数据案例：

BUYER LASTICO DE INGENIERIA S.A.

Date/ 日期 2010-01

Company/ 公司名称 PLASTICO DE INGENIERIA S.A.

HS Code/ 海关编码 39209200

HS Description/ 编码描述 OTHER PLATES，SHEETS AND STRIPSS OF POLYAMIDES

Details/ 详细描述 PLACAS DE POLIAMIDA|ROCHLING-F|50 X 620 X 3000|NO CELULAR SIN REFUERZO ESTRAT|IFICACION NISOPORTE SIN COMBIN|AR CON OTRAS MATERIAS|2141 KN

Customs/ 海关 VALPARAISO

Transport/ 运输方式 BY WATER

Country_origin/ 原产国 GERMANY

Country_shipped/ 启运国 GERMANY

CIF(US)/CIF 价 10386.2200

Quantity/ 统计数量 2141.0000

Unit/ 计量单位 KILOGRAMS

Net Weight/ 净重 10780.0000

Gross Weight/ 毛重 10780.0000

Loading Port/ 装货港 AMBERES

Unloading Port/ 卸货港 VALPARAISO

Freight(US)/ 运费 421.2000

Insurance(US)/ 保险费　195.3900

以下是一个提单的数据案例：

BUYER：TRI-S INTERNATIONAL INC. 20 ROYAL

Act_arrival_date/ 到港日期　2010-01

Est_arrival_date/ 预计抵达日期　2010-01-05

Shipper/ 发货人　TECNO ART MARMI SRL

Shipper_ADDR1/ 发货人地址 1　VIA NETTUNESE KM 13

Shipper_ADDR2/ 发货人地址 2　ARICCIA

Shipper_ADDR3/ 发货人地址 3

Shipper_ADDR4/ 发货人地址 4

Consignee/ 收货人　SHAW INDUSTRIES GROUP INC

Consignee_ADDR1/ 收货人地址 1　616 E WALNUT AVE

Consignee_ADDR2/ 收货人地址 2　DALTON GA 30722-2300

Consignee_ADDR3/ 收货人地址 3

Consignee_ADDR4/ 收货人地址 4

Notify/ 通知方　TRANS TRADE USA INC

Notify_ADDR1/ 通知方地址 1　1040 TRADE AVENUE

Notify_ADDR2/ 通知方地址 2　IRVING TX 75063

Notify_ADDR3/ 通知方地址 3

Notify_ADDR4/ 通知方地址 4

Container Number/ 集装箱号　UACU3171616

Piece Count/ 件数　49

Description/ 货物描述　TRAVERTINE TILES SLAC 49 PCS EXPRESS RELEASE E-MAIL DO W TRUCKER INFO TO UAAIDELVRYORDER@UASC.NET ALSO SEE GOAL108439

Carrier Code/ 承运人代码　UASU

Vessel Country Code/ 船东国家代码　AE

Vessel Name/ 船名	AL ABDALI
Voyage Number/ 航次	0033W
Bill Of Lading nbr/ 提单号	UASUGOAL108440
Foreign Port Lading/ 启运港代码	47531
Place Receipt/ 发货地	LA SPEZIA7d!Q-\8{ ‘}(l4^
Port Name/ 卸货港	SAVANNAH，GA.
Manifest Qty/ 载货数量	49
Manifest Units/ 载货单位	PCS
Weight/ 重量	111
Weight Unit/ 重量单位	KGt1J4T5G4J%k8P9N
Measurement/ 尺寸	25
Measurement Unit/ 尺寸单位	CM
Remarks/ 备注	PO. 790133 | NO MARKS | NO MARKS | NO MARKS | NO MARKS | NO MARKS

同行看到这些数据，可以得到很多有用的信息：从发货单位可以知道有哪些行业竞争对手、行业潜在客户或供应商。从装货港可以预测货物产地、竞争对手的分布。从数量可以分析同行的生产规模、供货能力、供货总量、市场份额。从卸货港可以预测最终消费地区、采购商分布。从收货单位可以了解老客户忠诚度、发现潜在客户、找机会夺回已失去客户。

从信息安全性角度看，如果想跟任何一个出口企业索取这些数据的话，他一定会以商业秘密为借口拒绝提供。

事实上，这些互相竞争的出口企业并没有由于商业秘密的泄露而破产，国际贸易也没有由于海关数据的公开而崩溃。不但不受影响，反而促进了国际贸易。

所以说，隐私不能成为阻碍数据开放的借口。

1.5 美国的启示

中国现阶段的主要目标是跨越中等收入陷阱，进入发达国家行列。考察作为世界最大的发达国家——美国，对中国未来的发展具有重要意义。虽然考察过美国的人很多，但仁者见仁，智者见智，不同人对同个事物和现象有不同着眼点和看法。

笔者分两次在美国的东、西海岸做了自驾游，游览的同时对美国的经济社会进行了一些观察和思考，并且平时也比较关注对美国的一些报道。儿子在美国留学期间也反馈了一些信息。

两次自驾游每次都是半个月时间。第一次是在西海岸，包括加州全境，从北边的旧金山一直到南边的圣地亚哥，往东到凤凰城，北边到大峡谷和拉斯维加斯。第二次主要是东北部，水牛城到波士顿、纽约、费城、华盛顿这条线。因为是自驾游，所以比较自由，住宿、租车、吃饭基本都是自己安排。在东部拜访了两个同学，到他们家里做客。通过考察和思考，我觉得美国有很多东西值得我们学习。

在美国主要有几个感受：第一，工资高，人工成本高，人员的成本费用高，从人员工资高可以解释很多的现象；第二，资源配置比较均衡，区域之间差异比较小；第三，相对收入来说美国的物价非常低。

很多人认为发达国家应该收入高，同时物价也高。实际上，美国相对于它的人均收入来说，其物价是非常低的。举个例子，不考虑币种和汇率，美国人可能一个月有 5 000 元的收入，中国人也有 5 000 元的收入，但在美国，苹果手机只是 600 多美元，而在中国，苹果手机却要 5 000 多元人民币，从这个比例来看，中国的物价和收入比将近美国的 8 倍。

那么，在美国的人力成本这么高，为什么美国的物价会这么低，它是如何做到的？

首先要搞清发达国家的真实含义。所谓的发达就是人均收入高，购买力强，生活水平高才叫发达。如果一个社会人均收入高，物价也高，显然

这不叫发达，而叫通货膨胀。只有保持物价低，才能体现出发达。

据作者观察分析，美国物价低的原因有两个：第一是公共成本低；第二是效率高。这里的公共成本是指所有企业和民众都会承担的成本，比如土地、电力、高速公路、汽油。效率高的主要特点是：

（1）自动化程度高；

（2）自助项目较多；

（3）关注流程设计。

这些实际上都是由于人力成本高导致的。在麦当劳刚到中国时，它的标准化产品和销售流程给中国人很多的震撼，但这种设计在美国实际上已经是一种标准化的设计。在中国由于人力资源比较丰富，这种对流程的设计不太讲究，所以才觉得新奇。

在加州从 1 号公路去丹麦村索尔文的路上，看到一个典型的加油站，加油站大概有 8 个加油柱，里面还有个较大的超市，而整个这样大的加油站只用了一个人。一个人怎么胜任工作的呢？外面的加油都是自助的，用信用卡可以在加油柱上直接刷卡，也可以到超市里面刷卡或付现金。付款都由这个人负责。告诉他你的汽车停在几号加油柱，付完款后就会开通几号加油柱加油。超市有个很大的窗户，他在里面完全可以看到外面加油站的加油情况。超市里有很多东西是自助的，比如说咖啡你可以付了钱以后，自助选择自己需要的咖啡种类。他一边观察店里的情况，一边收款，而一些小东西布置得很方便，比如要买香烟，他在头上柜子一伸手就可以拿到。

另外，在美国租住旅馆，在退房时是不会查房的。这种现象被国内很多人解释为美国人的信用度比较高。我的理解是主要查房的成本太高，即使丢了东西，在美国物价很便宜，其损失远远低于雇一个人要支付的成本。一家旅馆里，一个负责登记的服务员同时也兼管早餐服务。

在美国各地，购物中心都集中在一个叫 Plaza 的广场上。这个广场的店面配置非常科学，每个店都非常个性化，相互之间错位竞争。这些店都比较大，比如有综合性的百货商店如梅西百货，有专卖办公用品的 Staples(史泰博)，有专卖婴幼儿用品等，还有不同风格的快餐店。这些店用人很少，基

本都是连锁店。美国只有在唐人街和墨西哥人居住的地方有小店。

比如在洛杉矶，如果所有的商场必须建在市中心，那它的地价肯定会很高，如果可以选在任意的地方建商场，成本就会很低。毕竟如果可以任意选一个地方围一个四方形，主要中间建个停车场，周边就可以开店，地主就不能随便要价。

在美国，一般吃一顿西餐快餐在 8 美元左右，鸡蛋最便宜的 1 美元能买 12 个，当然这跟美元相对币值比较高有关，跟其他的货币相比它换回的石油价格比较低，但最关键的还是它的整个商品经营的成本比较低。

由于美国汽车普及率高，所以它的商店可以设在任意的地方，这样就可以避免类似中国高地价的困扰。当然在美国东部像纽约这样的城市，它的地价还是很高的。另外它的油价很便宜，高速公路很多都是免费的，西部基本是免费的，东部只有部分收费。美国只要是基础需求，就便宜甚至免费，因为越是基础的东西在物价中出现的概率越高。

可以想象，在中国商场里买件衣服，一般商场标价都是几千元，为什么会这么高呢？因为在这件衣服里面，切分这块蛋糕的人太多：第一，商场要分掉一块，大概要分掉四成；第二，租柜台的经销商要分掉一块；第三，生产厂家又分掉一块。生产厂家有物流费用，物流中是高速公路的垄断收费。生产成本又包括了在当地租房的费用，房租中是政府垄断的土地费用。

在美国，像梅西百货销售的商品都是自己直接采购的，它的利润就是从出厂价到零售价的差价，基本上就是一个公司在赚差价。而在中国多了个中间商。房租在成本里面占比很大，商场有很高的房租，生产厂家也有很高的房租，两重房租放在里面。美国则地价便宜，像梅西百货可以开在比较偏僻的 Plaza 里面，甚至是自己拥有的产权，地价几乎可以忽略不计。

美国的资源在各地配置比较均衡，即使在很偏僻的乡镇，也可以享受很高的生活质量。甚至越是偏的地方，生活质量越高，原因是汽车文化的发达。虽然地理位置很偏，但是非常容易开车到城市上班、购物。美国的地价很便宜，可以在任何一个地方建造广场开店。不管在什么地方都能买到类似的商品，同样品质、同样价格。

对美国的考察和思考的目的，是希望能为数据时代的发展指明方向。

一个社会要发展，必须是资源的配置均衡。实现均衡，其中一个是政治制度上的改革，比如说土地制度，还有公用事业的收费管制。本书有专门一章会讲到通过数据的公开和透明，加强社会对公用事业的监管，达到有效地降低基本收费的目的。

相对于其他成本，土地成本、高速公路收费增加的物流成本、电力、水力垄断成本，这种公共成本对最终商品成本的增加有一个累加效应，对物价的影响非常大。通过数据共享，加大信息的传播，使不同的地方都能得到相同的信息，使生产资源在全国得到均衡的分布，能够大大降低成本。

1.6 数据的价值与变现

1.6.1 数据的变现

大数据已成为新兴产业的热点之一，但也遇到很大的问题，就是大数据如何变现的问题。我们需要分析一下已知的数据变现的案例，才能找到变现的通用路径。

从沃尔玛的啤酒和尿布的故事，以及现在购物网站上的推荐，可以把数据的变现分成两个环节：一个环节为显示数据；另一个环节为决策。

在沃尔玛的尿布和啤酒的故事中，首先是通过数据挖掘发现啤酒的销售跟尿布销售的关联性，由此得到数据挖掘的一个数据结果；其次是管理人员根据数据结果做出决策，在超市货架上将啤酒摆放在尿布的旁边，因此增加销售，产生额外的效益。

同样，从购物网站的推荐的工作中也可以看出相似的过程。如果在网上点了一个尿布，网站推荐一个啤酒，购物者由此获得数据挖掘的结果数据，但如果购物者不做相应的决策，即不选择啤酒，而直接将尿布放到购物车，并最终完成付款，那么这个数据挖掘工作是没有价值的。

由此可见，大数据开发结果本身并不能产生直接效益，它通过影响管

理层的决策而产生间接效益。管理人员根据数据做出决策，正确的决策及相应的执行才产生价值。如同战争中正确的情报带来的胜利。

一个决策的影响，大到数十亿美元盈亏的投资，小到只浪费点汽油和时间的出行。无论政府还是企业、个人，无时无刻不在做出各种决策。每个决策都必须依赖足够的信息，而信息都来自数据。用数据产生的结果引导决策，可以产生直接的效益。

1.6.2 决策产生价值

数据通过为决策提供支持而间接产生价值，即人们是通过决策来实现数据的价值。决策离不开可靠的信息，数据是信息的主要来源，数据通过转换变为决策者可利用的信息而获得价值，并且得到回报。

决策可分为自动决策和人工决策。自动决策虽然更为直接和方便，但可应用场合较少，更多的为人工决策。所以数据技术的本质是将物理上产生的大数据转换成人眼可识别的小数据，再将小数据变为大脑可以快速直观吸取的信息，从而产生它的价值。

在互联网上采用数据挖掘就是典型的大数据应用。图书电子商务网站会通过搜集消费者以前的购物消费习惯，对消费者过去浏览过的、购买的书籍以及在购买其他商品的同时购买的书籍进行数据挖掘，一般采用的是购物篮分析算法。当一个新的用户登录网站后选择了一本书，网站后台工作程序就可通过有方向的数据挖掘得出相关书籍推荐，并且在快速计算后将相关书籍的清单展示给该用户，实现一对一的推荐。

但是，这种后台数据挖掘的计算以及书籍的推荐显然没有产生任何效益，只有用户对于网站的自动推荐产生兴趣，点击推荐的书并且加入购物车购买以后，整个流程才会增加效益、产生价值。

所以，真正变现的环节是人的选择，其他的只是参考。假设有一个人知道网站推荐是有目的的推荐，从而有意不点击推荐项目，那么数据挖掘的任何工作都不会产生效益。由此可以得出，大数据应用最后产生价值的主要环节在于人们的决策。

在 CCTV2 的一期财经节目中，主持人邀请京东 CEO 刘强东和财经作家吴晓波一起座谈，其中提到京东客户购买手机时从下单到快递员送货上门只需 7 分钟，其速度相当快。

在这个惊人的大数据应用案例中，京东通过大数据预测预先把手机派送到小区附近，当客户下单时实现了迅速送货上门。京东是通过对流程进行分析，用大数据对购买趋势进行预测，再根据预测结果，派送员将货物派送到小区。由此可见，这其中的核心还是派送员的决策。

作为一个大数据在电商行业的典型案例，其他行业难以直接模仿。这些行业中，遇到的问题和电商不同，数据的作用不在于如何快速送货，但应该可以找出这些行业在日常工作中的决策点，即需要决策的是什么，什么能够提高决策执行的速度，做这些决策时需要什么数据，能够提前做什么，从这些角度思考就能发现大数据的价值。所以尽管其他单位的工作场景在自己公司不能实现，但每个单位无时无刻不在做大大小小的决策。

所以，应该从决策的角度分析问题，找到很多大数据的应用领域。

1.6.3 数据的价值特点

作为决策支持工具的数据，如果把它当作一个产品，它和其他产品有什么区别呢？

第一大特征，数据的价值可有可无，可以被利用也可以被忽视。“可有可无”是指数据只是作为决策的辅助工具，当人在做决策的时候可以用到数据也可以不用。如同战争时期，不论有无情报都可以打仗，区别在于是打了胜仗还是打了败仗。没有情报有时也可以打胜仗。在图书电商网站买书，无论有没有推荐都可以买书，区别在买多还是买少。没有推荐有人也会买很多书。数据对决策的结果有影响，但对行为并非必须，这是第一大特征。

第二大特征，数量可多可少。数据越多，决策的正确性越高，胜算越大，但这些都不是必须的，很少的关键数据也能影响很大的决策。

第三个特征，价值可大可小。即利用数据后最小收益可能开车在路上

节约了 10 分钟，而最大收益可能是在一个投资项目中获得数十亿元人民币的收益，所以它的价值具有不确定性。

而其他产品，比如手机，拥有手机就可以在移动状态下打电话，没有就不可以，有和没有是两种完全不同的状态。

1.6.4 数据服务的商业模式

在数据时代，会出现与数据有关的新服务。这些服务主要集中在数据获取环节和数据的增值服务环节。

第一个环节为数据的采集和储存。这个工作主要负责采集数据，或者是负责从不同的数据源收集数据把它集中起来，或者将不可机读的数据转化为可机读。

第二个环节为数据增值服务。这个工作在拿到数据之后，提供依赖于数据的服务。比如，开发一个可以利用这些数据的软件系统，或者把数据和软件打包后面向最终用户提供云服务。也有可能只提供一个解决方案，而数据由客户自己购买或用客户自己的数据，最终客户直接将这些数据用于决策而不需要二次开发。

数据提供一般有三种模式：第一种是提供最终数据的查询，提供一个满足检索条件的数据集合，需要唯一的条件匹配，比如身份证号码、企业代码证号码；第二种是提供统计数据，根据查询条件给出统计数据，但不提供个体数据；第三种是提供原始粒度的数据，按照本书的介绍，如若要采用“鹰眼”技术，则必须采用原始粒度的数据进行分析。

每一个数据采集和服务商都不希望自己仅成为一个数据的提供者，而是希望提供更多的增值服务。但是，客户的需求多种多样，难以确定客户需要按什么维度去统计。对数据的汇总实际就是对数据维度的裁剪，就是对数据有效信息的过滤，仅提供统计数据会明显限制客户可以利用数据实现的功能，也减小了数据服务商的市场。

按照专业分工的要求，修路就是修路，开车就是开车，不可能哪家公司修了高速公路还必须租这家公司的车才能在上面走。同理，如果数据提

供商要求客户只能采购自己的软件访问数据，而自己的软件功能又不能满足客户需求，就会违反专业分工的要求，路会越走越窄。

实际上，卖数据是一个很好的商业模式。卖方提供数据，买方向卖方订购数据，因为数据是不断更新的，所以买方买的是旧数据，第二天又会购买新的数据，这种盈利模式没有问题。

数据服务商为保护自己的数据资源，最好的方式不是不允许别人把自己的数据装载到他的服务器上，而是在技术上提供更方便的模式，可以方便地直接访问放在云服务器上的数据，而不需要抽取数据。

根据数据仓库技术，需要在本地建立一个数据仓库（或数据集市）服务器，把原始数据从异构的数据源中抽取过来。如果原始数据不是关系数据库，可能会需要先建立一个关系数据库，将原始数据导入到这个数据库中，再通过编制的 ETL 程序把数据放入数据仓库里。

这种模式多了一个比较麻烦的环节。如果作为一个数据提供商，可以提供一个接口，只要编制一个 SQL 语句加上 IP 地址和一定格式参数，直接访问服务器，就可以定时提取数据。这样不仅提高了效率，也减少了客户本地服务器的存储，而数据提供商可以在客户订购期间提供数据访问服务，一旦合同期限到就终止数据访问，形成自己的商业模式。

1.7 信息时代遗留的问题

1.7.1 缺乏原始数据

如果去研究一下国内外出版物，特别是有关社会与经济发展的书籍，我们会发现一个共同点：只有结论而没有原始数据。

中国社会科学院每年发表一本《中国城市竞争力报告》，设计各种指标，从多个方面对城市竞争力进行排名。虽然研究的价值和影响很大，但遗憾的是未能提供研究结论的原始数据。虽然有些指标可以从其他数据来源计

算，比如，房价收入比、居民消费购物场所数以及人均住房面积等，但还有很多指标并没有来源。类似的还有很多研究区域经济发展情况的文章或书籍，可能只会给出同比增长率的数据，没有提供计算这些增长率的当期和同期数据。当然，限于文章或书籍的篇幅，提供全面的数据比较困难，但如果做一个规模较大及连续的研究，开发一个提供原始数据的网站可能对社会能提供更大的价值。

这样的社会科学研究，后人无法在这个基础上进一步研究，也无从确认研究结果的真实性和准确性。后人的研究就只能从简单的数据收集开始，做大量的重复工作。

《当代生物学》（*Current Biology*）在2013年12月发表的一篇论文中[4]，研究了1991—2011年的516篇文献，发现在论文发表20年之后，原始数据有80%丢失。由此看来，人们对原始数据的保存非常不到位。

假设几十年后，有人研究中国改革开放30年的经济发展，他除了这些提供间接数据的文章之外，将找不到任何可供研究的原始数据。理论上说，虽然我们现在身处于一个知识爆炸、信息发达的时代，但真正的实质数据还是相当缺乏。

历史学家研究古代历史只能靠发掘陵墓发现新的文物，从考古的重要性来说，文物肯定不如文字，比如在陵墓里面发现的甲骨文或者竹简，它上面的信息价值要大于文物本身。

后人要研究我们这个时代，不能仅仅依靠文字和图片。数据会和文物、书籍、绘画一样，成为记录一个时代的载体。所以数据的重要性显而易见，我们要注重保管好这些数据。

1.7.2　难搞的需求

阿基米德说过，只要“给我一个支点，我就可以撬动整个地球”。程序员也说，只要给我一个需求，我可以开发任何软件。

实际上，程序员说的需求，不是简单的“一个需求”，而是包括如何满足需求的设计。这个需求一般包括想要什么功能，操作流程如何，甚至

程序界面如何布局。很多软件公司和程序员以此作为骄傲，证明自己的技术很强。

如果在建筑行业，一个建筑工人可能会说，只要设计师能设计出房子我就能建造出来。这里的设计不仅仅是一个简单的需求，或者简单的设计概念图，肯定包括详细的尺寸、材料等。有了设计，对于建筑工人来说确实简单了，但设计才是困难所在。

因此，在信息系统开发中，需求是一个比较困难的问题。

在一般的事务处理软件开发中，一个软件实际上是把现实中由人工处理的流程转移到计算机上来，而人工处理流程常常是运作多年的成熟流程。在一个实施信息化初期的单位，可能只是希望用计算机实现原来手工处理的流程，所以它可以详细描述原来手工处理的流程，作为一个需求。

随着信息化水平的提高，很多客户不再满足自己提出需求，觉得自己的管理不够完善，希望借鉴别人比较先进的流程，学习其他领先公司的管理。因此，越来越多公司倾向购买现成的成熟软件。

实际上，中国人和日本人一样，还是比较喜欢完全按照自己的流程来定制软件，而美国人就不太一样，他们比较喜欢用现成的软件，这种偏好实际上反映了流程运作的规范化程度。

但是，在 DSS（决策支持系统）开发中是没有需求的。因为在日常的工作中，并没有决策支持的标准流程，毕竟需要决策的事件都是随机发生的，所以也没有标准的处理流程。

按照一般事务处理软件开发的逻辑，既然没有流程也就没有需求，没有需求也就无法描述给软件开发人员，开发人员也因此无法开发软件，即使开发出来也需要经常修改以满足客户不断变化的需求。

那么，如何在没有需求的情况下开发 DSS 系统呢？主要技术在于数据模型的使用和对数据源的研究。

数据模型是前人总结众多需求得出的，因此，一个客户的需求一般都能从模型中推导出来。

DSS 开发基础是现有的数据源。数据源来自客户的事务处理系统，事务处理系统包含了客户的需求。客户要看什么数据、不要看什么数据，在

数据源中都有体现。他需要看的数据，肯定已经录入数据库。如果没有这个数据，就说明他不需要看这个数据，否则需要先修改事务处理软件，增加该数据的录入功能。

科学研究方法有归纳法和演绎法。原来我们用的可能更多是归纳法，也就是从需求到软件开发，而现在用到演绎法，必须从模型去推导需求。

1.7.3 自助分析的陷阱

在商业智能（BI）领域里，2016 年发生了一个比较大的变化，由 Gartner 公司做的魔力象限把原来经典的品牌如 SAP、IBM、SAS 等都降到了有远见者的象限，而只在领导力象限留下了三个品牌，包括 Tableau 和 Qlik。

这个调整，把 BI 的方向导向了自助分析，如果说 BI 以前是由 IT 部门主导，现在则以业务部门主导。

但是，自助分析是不是未来的方向呢？笔者认为不是。

现在来分析一下自助分析的实质。举个吃饭的例子，有两种自助方式，一种是自助餐的形式，另外一种是 DIY 厨房式的形式。自助餐就是所有的菜都已经做好了，你只要拿筷子和勺子就可以直接吃；而 DIY 厨房式只提供厨房，需要自己去买菜做，但可以做出任何自己想吃的菜，加任何自己想加的调料。

显然，现在在 BI 里面的自助是第二种“DIY 厨房式”的自助。也就是说，它只是提供了一种工具，业务部门的人员可以用它做出很漂亮的图形，但数据必须自己处理，图形必须自己选择，需求也是根据自己的业务需要去设计。这种自助对人员的要求相对来说比较高，起码要熟悉自助分析的软件。虽然这种软件非常方便，也很直观，但毕竟需要学习。

但是，仅仅靠这个软件并不能解决 BI 所面临的问题，比如说大数据的问题，如何用很少时间从一个大的数据集中提取分类合计数据。像 Tableau 这种自助分析的工具，实际上只能面向一个有限的数据集，它的起点是打开一个数据平面文件，或者用一个 SQL 语言或 MDX 语言得到一个二维表，

也就是说，它提取的是一个有限的数据集，这个数据集要放在内存中，它的所有的分析都是在数据集的基础上进行分析。这个数据集开始是以一个二维表的方式给出的，虽然可以通过建立层次结构来变成三维或多维，但总的基础是一个二维表。

这种自助不是完全的自助，我们想拥有像自助餐一样的自助，就需要把“菜”完全做好，也就是说，把数据和需求都处理好，才能解决这个问题。

自助分析最终的目标应该是业务人员可以在自己的计算机上看到数据，而他的主要工作是去理解数据变化的含义，关注数据的趋势和不同指标之间的关系，从不同的维度观察同样的指标，分析它的同比、环比等。

总而言之，自助分析应该让业务人员不要去考虑数据如何获取，也不要考虑数据如何表现，重点是理解数据。而且，数据应该是来自一个很大的数据集，甚至是来自多元的、异构的数据集，但对使用者而言这些东西应该是透明的。

如果准确定义的话，Tableau 这类工具应该是自助分析的工具，自助分析系统是专门的人利用这种工具开发完成的。当然，做真正的自助分析时有这类工具也是非常方便的，毕竟，这让业务人员可以离开 IT 人员去完成比较简单的分析，应该来说是很大的进步。

Gartner 公司的魔力象限说明，现在国际上 BI 的发展方向和前沿技术是自助分析工具，至于真正的自助分析系统，则还有一定的距离。

自助分析系统实现虽然不能由业务人员独立实现，但可以为业务人员提供的方便性和强大的功能是自助分析工具所不具备的。

1.7.4　难以满足的客户

为什么客户总是不太满意定制开发的信息系统呢？

众所周知，当一个客户决定购买一个信息系统时，信息系统的供应商会给客户描述系统拥有的功能，这些功能除了满足一般事务处理的需求以外，更多的是描述完成以后对决策支持的帮助。

因为一般信息系统的采购是由高层管理者决定，而不是由底层业务人

员决定，所以高层管理者更多的关注信息系统完成以后对决策支持的作用，而信息系统供应商为投其所好，会从高层管理者的角度描述产品的功能。

但当信息系统实际完成投入使用以后，购买决策者会发现它的实际功能与供应商的描述及其自身的心理定位都有较大差距，从而产生不满。供应商为兑现自己的承诺，满足客户的需求，会按照客户需求开发一些定制报表，制作少部分装饰门面的统计图形。但这些程序运行速度比较慢，缺乏总体的一致性，其实是对客户的敷衍，不仅达不到客户的需求，还会花费大量的精力和成本。

问题产生的根本原因在于供需双方都把决策支持和事务处理混为一谈，低估了决策支持的难度。供应商由于技术限制，即使在知道决策支持难做的情况下，也只能从现有的技术出发进行开发。

事务系统和决策支持系统是两种不同的系统，不但开发的流程不同，使用的技术和工具软件也不同，甚至服务器都不能合用。好比买房子，建房和装潢是两个专业的事情，不能要求房屋开发公司同时也是一个非常好的装潢公司。

信息系统采购商和供应商之间的误解，可以用一个外国人来中国买房子的场景来比喻：一个对中国房地产市场一无所知的外国人想买房子，当房产商带他去参观样板房时，精致的样板房装修营造的舒适的生活环境打动了他，他会认为这就是他要买的房子，即所谓的精装修房，从而顺利签约。等到交房时，这个外国人发现却只是一个毛坯房，与他想象中的房子有很大差异，从而与房产开发商产生纠纷，要求房产开发商按照样板房交付。产生纠纷的原因可能是房产商当时并没有说清楚，也可能他没有理解房地产商口中的毛坯房和他看到的样板房有这么大的区别。

如果房产商决定迁就外国购房者的要求，按照他的要求进行精装修，但由于他们对装潢缺乏总体设计，即使能提供一些基础设备也达不到样板房的效果。

在建设信息系统时，供应商描述的功能类似样板房，等实际交付时用户发现却是毛坯房，从而产生纠纷。如果一个组织的负责人从来没有参与过信息系统的采购，他将难以分辨出哪些是供应商对前景的描述（样板房），

哪些是供应商实际能够提供的功能（毛坯房）。

总体而言，任何信息系统都是由事务处理和决策支持两部分组成。这就像买房子一样，一个是房产商提供，另一个是装修公司提供，必须由不同的供应商提供。而最关键的问题在于，现在一般的事务系统供应商并不掌握决策支持所用的技术，这两种技术是完全不同的。

如果一些事务处理系统能够和决策支持系统结合，那么对客户的而言能够提供更大的价值，并且可以避免客户验收的后期在报表上与他纠缠，避免造成花费大量的人力物力后还是不能满足客户需求的后果。

因为报表系统是二维的表，虽然用了比较复杂的嵌套表头，但它还是一种局限于二维的数据展示工具，另外由于它没有图形，很多时候合计速度比较慢，难以担当决策支持系统的重任。

1.7.5 完全不一样的需求

在计算机出现的早期，人类对计算机能够实现的功能只有一个模糊的需求。当一个计算机业内人士向圈外人描述计算机功能的时候，他会把可以随意浏览的数据和将数据可视化当成一种基本功能，但这种看似简单的需求在信息技术革命已经结束时仍然没有得以实现。人们最初认为，需求只有一类，并可以通过相同的技术来解决，但后来经专家研究后发现这种模糊的需求必须分为两类：一类是事务处理；另一类是决策支持。由于这两类需求对技术的要求不一样，所以需要采取不同的技术来解决。

经过几十年的发展以后，事务处理方面的需求得到很大的满足，这也导致了现在大量的计算机应用都局限于事务处理，而决策支持这块的需求由于适用技术缺乏而被忽视。20 世纪 90 年代才出现相关的数据仓库技术，并且在该技术基础上开发出大量的商业智能软件，包括 ETL、OLAP 以及图形展示工具。但是，商业智能软件在火爆一段时间后，在实际应用上并未获得市场的认可，导致现在这方面的应用仍然非常落后。

因此，计算机发明伊始，人类基于希望的两类功能迄今为止只满足了一个，另一个仍然缺失。根本原因在于，两类功能都把客户需求当作开发

的前提条件。实际上，在数据仓库的创始人比尔·恩门的书中[4]已经明确说到，这两类应用的最大区别是：事务处理是先有需求后有开发，而决策支持是先有开发后有需求。

虽然拥有了商业智能的技术和软件，但一方面，由于大多数从事商业智能软件开发的技术人员都是从事务处理软件开发转过来的，而且很多项目对这两类功能没有明确分割，所以技术人员存在固定思维习惯，必须要先有需求才能进行开发，非常不适应没有需求的开发。另一方面，虽然数据仓库的理论里有很多的技术可以应用（比如维度模型），但熟悉这方面技术的人非常少，没有人能找出进行无需求开发的方法，所以难以应对没有客户需求的开发。

只有充分利用数据仓库的维度模型，把维度模型的价值充分发挥出来，通过过度设计、模型推导来应对可能出现的各种各样需求，才可能实现无需求的开发。

1.7.6 心有余而力不足的数据挖掘

经过多年的努力，很多公司在商业智能的开发中开发出许多相关的产品，最后由大的 IT 公司，例如，IBM、SAP、微软、Oracle 进行收购整合，形成完整解决方案。

目前，商业智能软件或者 BI 软件开发和整合高潮已经过去，很少再看到新的技术出现。但 BI 软件的应用也没有取到预想的效果。从市场反馈的信息看，成功实施 BI 的公司很少。表面是由于价格昂贵而客户的需求少，实际上是实施失败率高才导致的价格昂贵，因此价格高只是一个结果而不是原因。

纵观整个商业智能软件体系，可以发现，软件开发将数据挖掘作为商业智能发挥作用的一个主要方向。即在整个商业智能技术方案架构中，ETL 或 OLAP 都是内部的一种技术实现，而展示和数据挖掘是最终向用户展示效果的主要手段。

但是，用户对 BI 展示和数据挖掘效果都不太认可。主要原因是数据

挖掘对需求的要求比较高，展示开发也是根据需求来进行的。一般数据挖掘属于有方向的数据挖掘，必须按照需求来进行，而这方面需求在决策支持系统的理论上，一开始就认定没有需求，从而为成功设定了高门槛。

对于用户来说，进行何种决策是不确定的，因此无法确定需求的具体内容和需求的使用频率。即使能有需求，也只能满足用户常见的、可重复的场景，这种场景对用户来说价值非常低。在用固定的算法进行数据挖掘时，对需求的限定范围更窄，相应可应用的场景也就更少。

从微软公司的开发历史来看，微软公司从 SQL Server 2000 开始支持商业智能功能，随后在 SQL Server 2005 开发出一套数据挖掘工具，实际上是对数据挖掘中的某些算法给予了支持，但在后来的版本中则对数据挖掘没有进一步开发。其原因很简单，它能提供的数据挖掘算法太少，不能满足客户的需求。数据挖掘成熟的算法很少，适应面也很狭隘，微软又仅仅实现其中很小的一部分算法，众多的限制使其可应用的场景非常少，因此微软停止了对数据挖掘软件包的开发。

现在数据挖掘方面发展最好的就是“R 语言”，由于它是一个开源的系统，故而很多人在这里提供数据挖掘的新算法并且可由他人进行修改，因此它带的数据挖掘软件包很多，算法资源非常丰富，适应性非常广。

由于数据挖掘算法的复杂性，没有哪家公司或者个人可以推导出一个有限的数据挖掘算法集合。就像苹果公司只能开发手机，而无法垄断 APP 开发，因为它无法预测什么样的 APP 受欢迎，因此它聪明地只提供开发环境和 App Store，而让其他的开发者去开发软件。同样，数据挖掘算法也和 APP 一样，“R 语言”实际上是一个开放的环境，任何人都可以写自己数据挖掘的算法给别人使用，其中不乏一些经典的算法非常实用，但这也会对用户提出更高的要求，有很多的参数需要设置。另外，也不乏有些人鱼目混珠，发表错误的数据挖掘算法，所以如果要用这个算法必须经过自己的检验，这样虽然提升了使用的门槛，但这确实是在现有的技术环境下数据挖掘最好的解决方案。

实际上，针对数据挖掘最好的方法，就是通过统计汇总之后将大数据

变成小数据，随后导出标准的格式（如 CSV 格式），之后通过“R 语言”建立数据挖掘的模型，从而输出图形，若这些图形可以变成页面和其他的功能一起调用，就能基本上满足客户的需求。

1.7.7 跳出事务处理的红海

信息技术革命经过几十年的发展，软件的数量越来越多，并且重要性也越来越大。人们普遍认为，以后的世界将是一个软件定义的世界，以后所有的技术和设备的大部分功能都将依赖软件来实现。比如说，智能手机里肯定含有很多硬件，但智能手机功能的强弱并不是由硬件决定的，而是由它上面运行的 APP 软件来决定的。

在人们看到这么多的软件，并认识到它的重要性以后，会产生一种错觉，认为软件已经非常丰富甚至过剩。它的品种如此之多，涉及面如此之广，如果说目前软件还有不足的话，肯定不会被人认可。如同一个人身处闹市的中心时，看到周围都是人，会产生一种错觉，以为满世界都是人，已经人满为患了。实际上，如果有机会乘直升机从人群中往上飞，开始高度低的时候，在视野中还是有很多人，但当升到一定高度，就会发现人都集中在市中心一块区域中，除了市中心以外，周边还有很大的空地，人非常少。同样，在软件行业，也有这样的一个感觉：如果自己想开发一个软件产品，就会发现相似的软件产品已经非常多，但如果跳出这个圈子，会发现大部分的软件其实都可以归为一类，叫事务处理软件。

实际上，除事务处理软件外，还有一类软件，叫决策支持系统，却很少能看到这类软件。在组织做信息系统规划的时候，大家常常把决策支持系统和事务处理软件混为一谈，搞得整体架构非常混乱，关系也非常复杂。事实上，决策支持系统和事务处理软件是一个双生关系，而不是互不相关的两种类型软件，即每一种事务处理软件的数据都需要相应的决策支持系统处理，并不会因为有了事务处理软件就不要决策支持系统了，或者有了决策支持系统就不要事务处理软件。

所以，一般组织需要做两个信息系统规划：一个是事务处理软件规划；

另一个是决策支持系统规划。决策支持系统是基于数据的一种规划，用于指导事务处理软件的开发。

从软件的分类来说，事务处理软件是一片红海，而决策支持系统是蓝海。现在一般软件公司开发事务处理软件，如果它仅仅掌握软件技术是远远不够了，还必须对业务流程非常熟悉，必须有开发经验才可能被用户认可，而现在掌握决策支持系统技术的人才相对比较少。



第 2 章

认识数据革命

大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。

国务院《促进大数据发展行动纲要》

2.1 认识数据

2.1.1 数据分类

这里的数据指的全样数据，而不是抽样数据。有的数据不能达到大数据的要求，仍然需要被处理、认知。

大数据是一种特定的数据，它所需要的处理技术和方法和一般数据不同。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

数据按格式分为结构化数据和非结构化数据。非结构化数据是大数据技术发展的一个重点，但越来越多的人把声音、图片、视频归于非结构化数据，增加了大数据的复杂性。非结构化数据应该只包括办公文档、文本、JSON、XML、HTML 等（也称为半结构化数据），而声音、图片、视频应该归于多媒体数据。非结构化数据划分比较粗放的原因可能是目前大数据技术仅实现了大数据的存储，还没有实现对大数据的分析，如果深入对大数据进行分析后，就会发现这些数据之间有比较大的区别。与其他数据相比，从多媒体数据中现在无法提取可以进行统计分析的特征。一种可能的前景是，通过深度学习，从多媒体数据中提取特征，保存到结构化数据中，同其他结构化数据一起分析，最后钻透至原始数据级别，播放或显示相应的多媒体数据文件内容。

从数据量来说，多媒体数据量最大，非结构化数据数据量次之，结构化数据量最小。从数据含金量来看，结构化数据的含金量最高，对组织的价值更大，非结构化数据次之。

数据可以按来源分为内部数据和外部数据，内部数据来源组织内部的信息系统，数据真实、及时、准确、详细，使用没有数据隐私问题，也无须费用。外部数据可以来自网上公开的数据，比如上市公司的定期报告，也可能购买自第三方数据提供商。

从数据的使用对象分，有内部用户和外部用户，对外部用户主要要保护好数据隐私。

对数据的应用基本上是以上分类的一种组合。

2.1.2　数据来源和存储

数据是信息技术的产物，人类只有计算机诞生后，才能如此方便、大量地生成数据。

数据的第一个来源是人们通过计算机及上面运行的软件输入的数据，比如在企业 ERP 或者 OA 等应用软件上输入的数据，在社交网站上输入的数据，这是人们有意识、有目的地输入数据。

数据的第二个来源是人们在操作软件时留下的痕迹，比如网站的日志。

数据的第三个来源是机器运行时自动产生的数据，比如物联网或者是 DCS 控制系统产生的数据。

数据存储方式分为在线存储、离线存储和近线存储。

在线存储是指存储设备和所存储的数据时刻保持“在线”状态，可供用户随意读取，满足计算平台对数据访问的速度要求。就像 PC 中常用的磁盘存储模式一样。一般在线存储设备为磁盘和磁盘阵列等，价格相对昂贵，但性能较好。

离线存储是对在线存储数据的备份，以防范可能发生的数据“灾难”。离线存储的数据不常被调用，一般也远离系统应用，所以人们用“离线”来生动地描述这种存储方式。

离线存储介质上的数据在读写时是顺序进行的。当需要读取数据时，需要把磁带卷到头，再进行定位。当需要对已写入的数据进行修改时，所有的数据都需要全部进行改写。因此，离线存储的访问速度慢、效率低。

离线存储的典型产品是磁带库，价格相对低廉。

近线存储主要定位于客户在线存储和离线存储之间的应用。指将那些并不是经常用到，或者说数据的访问量并不大的数据存放在性能较低的存储设备上。但同时对这些的设备要求是寻址迅速、传输率高（例如对客户一些长期保存的不常用的文件的归档）。因此，相对来说近线存储对性能要求并不高，但要求相对较好的访问性能。同时，多数情况下由于不常用的数据要占总数据量的比较大的比重，这也就要求近线存储设备需要容量相对较大。

数据分为热数据、温数据和冷数据。热数据是指在事务处理系统中被频繁访问的数据，一般存储在快速存储器中。温数据被访问频率相对较低，一般存储在相对较慢的存储器中。冷数据指极少被访问的数据，会被存储在企业最慢的存储器中。

被备份的数据都是冷数据。随着计算机技术的发展，用于数据备份的存储器出现过各种各样的格式。最初在 IBM 的 PC 上用软盘备份，容量仅为 360KB，之后达到 1.2MB。一些大中型计算机采用磁带备份。后来出现了 650MB 的光盘，以及容量为 4GB 的 DVD 光盘。目前的存储设备主要有 U 盘及外接硬盘，也出现过刻录光盘，不过现在已经很少使用了。以上存储设备可以理解为个人存储，特点为数据量小，数据分割存储，适合数据量比较小的情况，但作为大数据来源的许多数据源可能就保存在这些设备上。

如今又出现了云存储，即处于热备份或者是温备份的状态，可以随时使用。当然云存储和数据的分布式存储还有很多问题需要解决，毕竟我们现在对数据的要求还处于一种温备份的状态，而且它分布在不同的机器上，数据来源不同且放在一起也有危险。分布式的存储目标以云中心为基础开发关于数据存储和数据查询的一些功能，并且形成分布式的存储管理方案，也就是说它是基于“云”但是不止一个“云”。这样，数据可以进行同时查询。有了“云”作为数据存储的目标，以前个体存储的方式都会被淘汰。

大数据的工作之一，就是把分散的数据综合在一起。把冷数据变成热数据，把离线存储变为在线存储。

2.1.3 非结构化数据

数据一般分成结构化数据和非结构化数据。现在大家比较关注的非结构化数据，比如，视频、图像、声音，实际上应该归为多媒体数据，标准的非结构化数据应该是不能通过普通关系数据库存取的文本数据，比如，JSON 格式或者 XML 格式数据。在数据时代，应该以结构化数据和标准的非结构化的数据为主，把非结构化和结构化数据一起处理。

至于视频、图像和声音，在数据时代它的作用不是很大，很可能属于下一个时代，而不属于数据时代。因为必须有技术从这些文档里提取可以保存的结构化数据，才可以对它进行利用，否则这种多媒体格式的文档作用不大。虽然从 Hadoop 的应用情况来看，存储多媒体数据成为一个主要目标，但它可能会误导发展方向，因为多媒体数据虽然数据量很大，但含金量很小。就像不一定是铁矿石就可以炼铁，钢铁厂会尽量采用含量高的铁矿石。视频、声音和图片的含金量目前来说是非常有限的。特别是视频，它占据的空间非常大，但含金量比较小，起码我们还没有能力发掘其中的价值。我们的精力应该主要放在这种结构化数据和不是多媒体的非结构化数据上。

2.1.4 数据处理的三个层次：产生、获取和分析

数据处理分数据的产生、数据的获取和数据的分析三个层次。

数据的产生在信息技术革命完成的今天已不成问题。现在不但存在大量的数据，同时随着物联网技术和应用的发展，已经出现爆发性增长，面临数据过剩的问题。在数据时代，技术开发的重点不应该在数据的获取上，它应该是信息技术发展的必然结果，而且，数据开发得越好，数据的产生会越多。虽然还有很多数据没有量化，比如有人在做量化历史工作，把历史上存在的档案录入计算机变成数据，但这只是一些特定领域的拾遗补阙工作，不是数据时代工作的主流。

其次是数据的获取问题。由于数据分布的离散性和格式的多样性，现在并不能很方便地获取数据。比如一个企业老板，按理来说他是企业最高

负责人，有权力获得企业的任意数据，并不存在安全和隐私问题，也没有资金的问题，但他仍然很难获取这些数据，一个主要原因是数据分布在不同的系统中。

信息化基础较好的公司，会有很多的信息系统，而每个信息系统都由不同的公司提供，并且放在不同的数据库系统中，不同系统的数据库格式有不同的定义。理论上，公司老板应该能够看见所有数据，但实际上他并不能获取完整数据，只能根据软件公司提供的程序查询部分数据。程序使用也不太方便，不同信息系统有不同的查询程序，同一个查询程序查询不同数据要调用不同功能。

放眼数据时代，任何一个人应该都能处理分布在全国不同单位的数据，但由于异构的数据库和不同的服务器，要整合数据涉及很多复杂问题，需要从技术上和法规上进行研究，开发出一种共享技术和机制。由于互联网不提供结构性数据而且对于权限的控制比较薄弱，更没有数据的整合，原来基于互联网的存取方式不可套用。

最后是数据的分析问题，即如何看懂数据。基于目前获取数据的困难，想要分析数据并得出有用结果更是难上加难。由于人类能力有限，对数据的分析需要把大数据转化为人类可以认知的小数据并可视化。

综上所述，数据时代的应用需要解决以上三个层次问题，其核心是数据的获取和分析。只有解决了这三个问题，数据时代才算真正到来。

2.1.5 数据比图像、视频更有价值

传统上人类认识客观世界的方法以“眼见为实”。按照信息技术的分类，实际上是相信图像和视频的信息。但是，图像和视频信息只能映射出事物的外表和某个时间点的静态信息，而无法看到事物内部的隐藏信息和时间历史信息。由于无法获取事物过去的图像和信息，仅靠“眼见为实”来获取静态图像和动态视频信息，如果没有其他的测量手段补充，这样识别事物是有局限性的，难以掌握其本质。

信息包括数据都是认识一个事物的必要补充，但人们总感觉数据所能

表达的信息很少，而图像可以表达很多。实际上，面对描述很长历史和众多个体的大数据，人类对数据的认识是非常有限的。随着个体的增多，单一个体的识别已失去价值，对群体的识别更为重要。此外，历史数据也应比单一时间点更为重要，但人类对把这两点结合起来的数据的应用还远远不够。

尽管现在人们对图像和视频的存储和识别花费很大精力，但由于占据存储空间很大，在硬件投资上占据很大的预算，从数据含金量的角度来看，他们所含的有价值的信息非常少。由于数据的含金量要大于图像和视频，因此，技术的重点应该放在数据的利用而不是图像和视频的利用上。

那么，在数据时代，图像和视频是否就没有价值了？当然不是。随着识别技术的发展，将可以从图像和视频中抽取有效信息，转换成数据格式，然后和其他数据一起被加以应用。

换言之，无论是现在还是未来，数据的利用才是最大的挑战，图像和视频最终作为大数据里面最小粒度的数据，只是在从一个很大的数据集中发现问题以后，需要钻透至最终的数据时，才会用到图像或者视频。

假设未来能够发明一种癌症识别技术，可以在一个人的照片中提取信息识别出该人是否是癌症患者，那么如何让这项技术发挥最大效益呢？难道仅仅靠医院提供癌症识别服务，让病人拿照片来逐一自动扫描，做出诊断吗？显然，通过和大数据技术的结合，还有更多价值可以挖掘。

具体做法是，通过收集大量癌症患者的图像，从中抽取出每个人的特征数据，建立数据库。利用数据分析，找出这些特征与不同癌症种类、发展阶段、治疗效果之间的关联。有了这些关联数据，就容易找出易患人群的特征，优先筛选的对象。对不同的治疗方案、治疗效果和存活率进行跟踪反馈。

什么时候需要用到个体的图像呢？在跟踪具体患者对象，开始进行治疗时。所以，图像变成数据钻取至最后得到的数据。

2.1.6 数据与程序要分离

一个独立的信息系统由硬件、软件以及数据库组成，计算机软件是由程序和相关的文档组成，数据由计算机软件产生，保存在数据库中。

现在一个软件项目，一般数据库和程序由同一家公司开发，软件由该公司交付，以后若由于该公司的倒闭等原因不能提供服务的话，该软件就会作废。

不同于软件设计与开发，建筑设计和施工是分开来的，先有设计图然后施工。建筑的设计师和施工人员需要完全不同的专业技术，由不同的公司来负责。比较而言，软件的设计和开发混为一体，甚至可由一个人来做，这种做法显然不合理。

一个比较合理的方式是将软件的设计和开发分开。软件设计工作包括数据库的设计和程序界面的设计，核心是数据库的设计。对于软件的设计，需要丰富的经验和对客户需求的准确把握，并且留有一定的扩展空间。一个不好的设计不符合数据库的设计原则，设计出来的数据库缺乏伸缩性，只能满足一时的需求，如果需求变动以后，修改需要很大工作量。

由于是同一个公司或个人兼设计和开发，不需要交流，因此很多数据库的设计缺乏必要的文档。如果软件被弃用，数据库的数据也被废弃，造成数据的浪费，影响数据的后续发掘利用。

理想做法是，首先数据库应该由经验相对丰富的人去设计，然后交给经验不是很丰富的人去编程。如果编程的公司不能维护，由于数据库文档还在，数据库的资源还能使用，可以找另外一家公司维护，只要有源代码，维护起来相对也比较方便。即使需要重编软件，由于数据库还在，和原来的软件兼容性也会比较好，更重要的是数据资源可以得到比较好的利用。

在关系数据库发明之前，程序和数据混在一起，彼此有很强的依赖性。但关系数据库出现后，由于它具有结构化程度高、冗余度低、独立性强等优点，从技术上支持实现软件中程度和数据的分离。

所以在项目开发中，程序和数据应该要分开。这种程序和数据分开的做法现在有些地方政府招标中已经开始实施，做法是先进行数据库的招标，再进行软件的招标。

2.1.7 SQL是访问数据的通用语言

结构化数据是以数据库或平面文件格式来存储的，可以用二维表的格

式描述，由行和列组成。一般列数固定，行数可变，可能有几万行、几十万行数据。列可以分为文字格式、日期格式、数字格式。一个最简单的二维表可以用文本格式来保存，文本格式一般称为 CSV 文件。CSV 文件每行用换行符分开来，行里面的字段可以用逗号或者制表键分开。在记事本等文本编辑器中看，每行的长度不一样，也不是对齐的，但如果用 Excel 程序打开的话，会发现已经自动对齐，一格一格的。CSV 是最简单的数据文件，一个文件里保存了一张表。在稍微复杂的 Excel 文件中，一个文件可以通过不同的工作表（Sheet）保存多张二维表。

更常用的数据存储模式是关系数据库。数据库里面包含很多二维表，称为数据表。数据库管理系统有简单的 Access，或者常用的 SQL Server，它们都由微软公司提供，还有大型企业用的 Oracle 数据库，其他还有 IBM 的 DB2、SAP 的 Hana 等。

一个数据库系统中可以包含多个数据库，一个数据库又有多个数据表。数据库不仅可以存在一个服务器上，也可以分布在网络中多台服务器上。

多个服务器可以通过网络相连，不同服务器上可以安装不同的数据库系统，通过任何一台连接到该网络的机器可以访问这些数据库。就是说，如果需要，用网络中一台机器上的数据库访问程序可以打开分布在不同的服务器上的数据库，即使服务器在国外，也只要联网即可访问。

还有些数据作为备份数据。一般备份数据是把数据保存在磁带上，这种数据平时不好打开访问，必须把数据恢复到数据库中才能看，因为它要占不少的空间，所以看完以后需要把数据删除，以便恢复和查询其他数据。

随着数据越来越多，出现了专门的数据仓库技术、数据仓库服务器。美国的 Teradata 公司专门做这种数据仓库服务器。数据仓库中的数据一般只增加，不删除和修改，只用于查询，保证查询时即使数据量很大访问速度也很快。

面对越来越大的数据库，一种处理方法是把一部分不用的数据备份起来，比如说三年以前的数据备份起来，不放数据库里面，平时也查不到，要查必须临时恢复。第二种处理方法是不断地扩充服务器，原来是一台数据服务器，现在要用 3 台或者 5 台甚至更多。

用扩充服务器的方式，不但要为增加的服务器付费，还要为相应的数据库系统软件付费。因为这种服务器是专用的服务器，软件许可跟服务器的数量有关，所以整个扩充一台服务器在软件和硬件上都有较大花费，一般只有 IT 预算比较宽裕的企业才这样做。

但是，即使愿意投入，这种数据库的扩充也有个限度，至于能否有个无边界的扩充，比如用普通 PC 服务器，用网络把几千台甚至几万台服务器连接起来存储数据，这就是要用大数据技术解决的问题了。

有一种叫 Hadoop 的系统专门用于解决这个问题。它的思路是只要用普通的计算机联网，就可以管理计算机中的数据，不管你的数据多大，只要增加服务器数量就可以。而且系统可以弹性配置，就是说你的数据满了，可以增加一台服务器，系统自动调整；有一台服务器坏了，可以把这台机器关掉，它自己脱离系统，存在这台服务器上的数据自动重新分配到其他服务器上。

Hadoop 中可以保存任意格式文件，比如视频。数据也是当作文件来保存的，而不像数据库系统专门处理数据。数据一般用最初介绍的 CSV 格式保存，然后通过一种名为 Hive 软件访问，模仿数据库访问形式读取数据。从外面看起来，好像里面存的也是类似数据库的二维表。

数据的检索有非常成熟的语言，叫 SQL 语言，或者叫结构化查询语言。这个语言已经非常成熟，通过语言的组合可以任意地查询数据库中的数据，并且不管有什么要求，都可以检索出来。SQL 语言有标准，也有不同的数据库公司对标准进行扩充，所以它的主要功能都是相同的。只要学会一种 SQL，不管在什么数据库系统上，基本上都可以运行，但不同的数据库系统有一些微妙的区别，比如查询两个字段并合并为一个文本显示出来，在 SQL Server 中用“+”就可以，在 Oracle 中用两个“||”才行。

现在很多跟数据打交道多的单位都加强了对 SQL 语言的培训。我国的审计单位很多的审计员都已熟悉掌握了 SQL 语言，这样他们就可以脱离程序对任何数据库进行检索。

为把存储在异地服务器上不同数据库系统中的数据合并查询，要建立分布式的数据仓库系统，并且以 SQL 语言为数据检索的基础。现在在网络

上检索文章，每次检索只能得到一篇文章，如果想把两篇文章合成一篇，可以分别查询，再利用软件进行编辑合并。在分布式数据仓库系统里面，近期目标可以先实现对检索数据导出后再进行合并处理；远期的目标一定可以进行联合查询，把处于多个服务器中的数据进行合并，直接输出一个数据集。

2.1.8 需要标准并开源的数据库设计

美国的软件产业比较发达，而日本相对来说比较落后，核心原因是效率。

美国软件有两种提供方式：一种方式是软件产品，即开发的通用商业软件，这个软件可以用于不同的单位，软件产品可以直接销售或提供服务（Saas）；另一种方式就是开源软件，当开发者觉得这个软件不成熟，自己无力独自完善和推广时，就把源代码开源，其他人可以在开源代码基础上继续完善，定制开发自己的软件。这两种模式导致的结果是：任何一个软件工程师的工作成果可以得到最大限度的应用，从而提高了整个社会的软件生产率。

日本公司是定制化开发，软件都是为特定用户开发的。一个软件工程师的工作成果只能被一个用户内部使用，不能为社会所共享，换一个用户就需要重新开发，因此日本整个软件生产率就比较低。

但现在美国的开源软件只是源代码的开源，还没有涉及数据库的开源。

在数据时代，更重要的是对数据库资源的开源和标准化。从软件开发角度看，数据库的定义常常能决定软件开发的效率和面向应用的弹性，也就是说，一个客户的需求主要体现在数据库的设计和用户界面的设计上，而这两个方面常常是比较关联的。数据库怎么设计，界面常常必须与之对应。

比如，数据库中常用的主从表的设计，一般有一个主表，记录一个订单什么时候下的，客户是谁，一个从表记录销售的明细，订单产品的规格、单价、总价。如果数据库里定义了主从表的结构，程序界面上就要支持这种主从表的结构。所以软件开发中数据库设计是一个非常重要的方面，特别是应用软件开发。

美国更多的是从事系统软件开发，所以数据库不是很重要。但作为应用软件开发，数据库则非常重要。数据库设计标准的开源和数据库标准的推荐是政府、学术团体可以做的工作，可以组织编一些比较标准的通用软件的数据库作为指导规范，提供给软件公司，软件公司再根据需求进行扩充。可以开发一个数据库的开源网站，首先把不同行业数据库设计的定义公开，其次就是进行推荐，这样可以大大提高社会劳动生产率。

在某些行业领域，比如说财务软件，数据交换已经有了国家标准。

开源数据库设计要求，首先满足功能要求；其次是有共性；最后是有弹性。还可以按应用规模，分成小型、中型、大型三个层次。小型数据结构比较简单，大型数据结构比较复杂一点。如此通过知识的扩散，能够大大提高社会软件开发的效率，

有了标准并开源的数据库结构以后，在数据时代，对数据的应用也会比较方便，因为大多数的数据库定义是基于开源结构，数据分析人员不需要花费太多时间就可以弄懂数据的含义，再提取到数据仓库中。

2.2 关于数据

2.2.1 数据和信息的区别

数据是一种客观存在，比如指定时间指定股票的成交价格；而信息是对数据的解读，比如该股票的价格是高还是低，该买进还是卖出，这些信息单靠成交价格无法判断，与环境有关，需要根据以前的价格和对未来预测及投资人的投资策略得出。

信息是对数据通过比较、分析、判断从而得出的结论。同一个数据在不同的时间会产生不同的信息，不同的人分析同一个数据也会产生不同的信息，不同区域的人对同一个数据进行分析同样会产生不同的信息。因此，信息和数据具有差异性，同样的数据会因外在环境和内在因素的变化而产

生不同的信息，而数据是客观存在的原始素材，不会随着时间的变迁和外部条件的变化而改变。

再举一个猪肉价格的例子说明同一个数据产生的不同信息。

商务部全国农产品商务信息公共服务平台发布 2016 年 7 月 25 日南京农副产品物流中心猪肉（白条猪）价格为每千克 22.4 元。如果在 2016 年 3 月看这个数据，得到的信息是猪肉价格涨了，不能买，因为 3 月的价格是每千克 16.48 元。如果是被位于常州的江苏凌家塘农副产品批发市场的人看到这个数据，得到的信息是这个价格比较低，可以买，因为这个市场当天的价格是每千克 26.5 元。

目前很多的书籍和文章给我们传递的信息，虽然在目前有较大价值，但随着时间的推移，会逐渐失去其原有的价值。由于书籍或文章中没有保存原始数据，后人将难以判别信息的真伪，也不能通过和新数据的比较生成新的信息。

由此可看出，数据时代应该保存的是数据而不是信息。

2.2.2 数据含金量

在资源领域有含量的概念。比如铁矿石，不同的矿山出品的铁矿石的含铁量不一样，不同的铁矿石炼出来的钢铁的出铁产量也不一样。为什么中国的钢铁厂现在都大量进口巴西和澳大利亚的铁矿石，就是因为这些地区的铁矿测出的铁含量非常高。中国的矿山开采出来的铁矿含铁量非常低，在可以选择含铁量比较高的矿石时，一般就会弃用这些含铁量比较低的。

同样，我们获取的数据也有含金量问题。如果按照字节去计算的话，虽然有些数据量很大，但它的含金量比较低。因此不是数据越大，价值越高。大数据的一个主要特征就是价值密度低。

在数据时代，不仅要关注数据，也要关注数据的含金量，要把更多的注意力投入到含金量比较高的数据上，而不是在含金量比较低的数据上不断投资。

如何区分不同数据的含金量？

首先，要从数据的格式上去区分。一般而言，结构化的数据含金量最高，声音、图像、视频等多媒体数据的含金量最低，非结构化数据介于中间。实际上，物联网数据的含金量也比较低，因为大量的数据都是相同的，它的价值主要表现在数据的变化上。

其次，从内容上来划分。一般交易数据的含金量最高，次之为日志数据，更差的为社交媒体的数据。

总而言之，从数据的开发应用角度来说，应该从含金量比较高的数据入手，然后逐步涉及含金量比较低的数据。

2.2.3 用于理解大数据的小数据

小数据不是指数量比较少的数据，也不是指用于描述细节的数据，更不是一个大数据通过检索条件过滤后的部分数据。

这里的小数据与大数据有密切的关联，它不是大数据的一部分，而是能够描述大数据特征的数据。

维基百科对小数据（Small data）的解释是：

小数据是能为人类理解的数量足够小的数据。是一种从容量和格式上都便于访问和操纵，含有用信息的数据。

“大数据”这个术语是关于机器的，而“小数据”是关于人的。这是说，可以一眼看清或比如只有五个相关数字的就是小数据。小数据是我们以前认为的数据。大约四分之一的人类大脑参与视觉处理。理解大数据的唯一方法是将数据变成小的、视觉上有吸引力的对象，这个对象能够表达大数据集的不同方面或具有被人类理解的“特征”（如直方图描述数据预测和关系、图表、散点图）。所以有时大数据被简化得像小数据。

2.2.4 广义和狭义大数据技术

虽然大数据的概念很火，但实际上对大数据技术的定义还是有些歧义的。

大数据技术可分为狭义大数据技术和广义大数据技术。

所谓的广义大数据技术就是包括BI在内的传统的决策支持系统以及为实现决策支持系统开发的一些商业智能工具，包括报表工具等。在涂子沛的《大数据革命》一书中，有一章专门讲到的大数据技术实际上就是商业智能技术。

从硅谷的技术人员的角度来看，大数据技术主要是指以Hadoop为主的一批开源的数据工具，而不包括传统的商业智能技术。这可以从开源软件中的菜单设置看出来。在商业智能软件中，有一类工具软件称为ETL软件，Kettle是一个开源的ETL软件，使用的人很多。若是从广义大数据角度来说，ETL工具本身就是大数据工具的一部分，但实际上，它的菜单中有一组功能，挂在Big Data的菜单下，都是针对Hadoop、HBase等开源软件的控件。即从Kettle的开发者的角度看，只有这些开源的软件才属于大数据技术，而Kettle软件本身不属于大数据技术。因此，可以把以Hadoop为主的软件称为狭义大数据技术。

从实际的技术发展来看，狭义大数据技术正是在商业智能技术处于发展停滞阶段以后推出来的一些新技术，他们的目标是一致的，应该也可以融合。比如Hadoop的出现实际上就解决了大数据原来用传统的数据仓库技术需要很大投资才能解决的问题。

随着狭义大数据技术的发展，出现了和BI技术结合的需求，比如eBay公司推出的Kylin（麒麟）开源系统，就是BI中的OLAP技术和Hadoop的结合。

OLAP有MOLAP、ROLAP和HOLAP之分。一般MOLAP和数据库是比较密切结合的，比如说微软的SSAS和SQL Server的结合，Oracle OLAP Server和Oracle数据库紧密结合，Kylin软件目标是实现和Hadoop的结合。ROLAP的技术并不需要把具体的数据保存在特定的数据库中，它只提供了一个访问接口，这个数据库完全可以把数据存在Hadoop中，通过Hive的接口来读取，这样仍然可以用原来BI提供的一些工具和接口去访问数据，只不过这些数据不是存在原来BI典型的关系数据库中，而是存在以Hadoop为核心的分布式文件系统中。

2.2.5 看懂数据的认知计算

很多领域需要用到数据。现在不是没有数据，而是怎么样去利用数据。IBM 把认知计算确立为转型后的重要战略支柱。“认知”这个名字笔者觉得还是非常适合的。数据不是有了就行，关键是要从数据中看到东西，要把数据包含的信息转换成能够识别、能够了解的信息才有价值。

IBM 定义的认知计算指的是要通过以人的自然语言交流及不断学习，通过技术与多个学术领域的结合使人们更好地从海量复杂的数据中获得更多洞察，从而做出更为精准的决策。

这里讲的认知不是计算机的认知，而是人类的认知，计算机只是帮助人类认知的工具。人类无法通过大数据认知，只能通过小数据。大数据只有转换为小数据，人们才能理解，才能认知。数据技术要解决如何把大数据转换成人能接收的小数据。

IBM 的认知计算包含领域比较多，把认知计算讲得过于神秘，和自己的优势技术结合得过于密切，把门槛提得很高，比如说利用沃森系统的自然语言的识别。认知不一定要自然语言，它的核心本质是对数据的认知，任何方式都可以，而且也没有一个非常具体的认知路线。虽然 IBM 的认知计算更多是一种战略层面包装，但这种思路或是提的角度是完全正确的，也可以消除现在物联网或大数据应用上的一些误区。

2.2.6 数据的冷态、温态和热态

借用冷数据、热数据和冷备份、热备份的概念，可以把一个组织内部的数据存在状态分为冷状态、温状态和热状态，分别称为冷态数据、温态数据和热态数据。

冷态数据是指数据处于不可访问状态，对应冷数据，原因可能是离线、不共享、无文档等状态。离线指数据保存在软盘、光盘、U 盘等媒介中，一般不好直接访问。不共享指数据存在的计算机虽然处于连线状态，但数据所在的文档不可以被其他计算机访问，比如一个单位中许多人都在自己桌面计算机中生成和保存 Excel 文件，文件所在目录和文件本身没有设置共

享属性，所以不能被其他人访问。无文档指虽然数据连续且共享，但缺少数据库、数据表、数据文件的目录、名称及内容说明，其他人找不到相应数据表或文件，或者找到了也读不懂。冷态数据对应目前大多数组织的数据现状。

温态数据是指数据可以被专业人员访问和利用，对应热数据或温数据。最简单的温态数据将所有数据文件放到一个文件服务器的共享目录中，并编制一个文件目录。典型的温态数据由多个关系型数据库和共享的多个平面数据文件目录组成，每个目录中包含许多 Excel 和 CSV 数据文件，熟悉 SQL 语言的 IT 或数据分析人员可以编制 SQL 语言任意访问数据库。高级的温态数据是数据仓库，比如银行用 Teradata 服务器和金融服务逻辑数据模型 (FSLDM) 建立的数据仓库，互联网公司用 Hadoop 集群建立的数据仓库。目前在大数据应用比较领先的公司中，把数据从冷态转换到温态，由专门的 IT 部门人员负责。

热态数据是指数据可以被决策人员和业务人员利用，比如传统报表、用 BI 展示工具制作的仪表板（Dashboard），使用者不需要使用 SQL 语言，不需要了解数据库或数据文件的位置、结构、内容，一般拥有以下特征。

（1）显示支持钻取的统计数据（如合计、平均值）。

（2）计算速度快，保证及时响应。

（3）所有数据可视化。可视化不是热态数据的唯一特征，它是建立在数据聚合基础上，需要用到数据集市和 OLAP 技术。

目前在大数据应用比较领先的公司中，把数据从温态转换到热态工作由专门的数据分析部门人员负责。

2.3 走出大数据应用误区

2.3.1 从个性化需求到普遍服务

目前的大数据应用中，大都要求结合用户的实际情况实现比较具体的

目标。如果一个企业希望使用大数据，它不但希望你能够提供数据、提供算法进行相应的开发，最终还要有一个比较明确的目标，对企业的效益要有明显的提升。如果这个工作对企业效益没有直接提升，它对这个项目就不会感兴趣。这种工作实际上是要求提供一种个性化服务，也就是做定制开发。

在任何一个产业的初期，都是被要求提供一些个性化的服务。比如在铁路运输开展业务初期，如果向用户推荐铁路运输服务，不考虑成本，他的需求肯定是点对点的服务，即无论是人还是货物必须从起点一直运到终点。实际上，我们现在的铁路服务只能从火车站到火车站，也就是说从起点到火车站、从终点到目的地这两部分必须利用其他交通工具来解决，铁路并没有提供解决方案。

理论上，铁路运输可以提供两种服务，一种服务是普遍服务，指火车站到火车站的交通；还有一种是个性化服务，即点对点服务。不过现在火车没有提供点对点服务，飞机提供的商务包机服务可以点对点，不但把机场到机场的交通解决了，而且从起点到机场、从机场到终点的服务也包含在内，但是这种服务收费较高，不是一般人可以接受的，因此这是一种小众的个性化服务。为大众提供的普遍服务都是机场到机场或是火车站到火车站的服务。

同样在数据应用上，今天的用户都提出了个性化服务，从而证明了数据的应用还处于早期。因此，市场更需要一种普遍服务，这种服务不一定能满足用户的直接需求。比如说不可能保证他有直接经济效益，但可以改进业务流程的某个重要环节。

所以，大数据的开发应该走出个性化服务的歧途，更多的应该探索一种普遍服务。普遍服务的特点是只能解决部分问题，要说服用户能够接受这种观点。用户接受这个观点应该也不是很困难，毕竟这种普遍服务的成本肯定比个性化服务要低。

2.3.2 走出结果导向

在目前大数据的应用中，主流的思路是有方向数据挖掘，因此有些用户过于强调结果导向，也就是追求有一个明确的目标，并根据目标去采集

数据，再为这些数据购置相应的硬件和软件，或专门建模分析。从表面上来看这种方式比较实用，实际上，是缺乏计划和远见。

以市政建设为例，看看结果导向会产生怎样的后果。如果我们现在要铺设排污管道，可能需要挖公路。挖公路这一工作的目标很明确，因为要铺设排水管道。如果下次要埋设天然气管道，那么又要挖公路。这就是原来我们经常说的马路上要安装拉链的笑话。为什么一开始不能建一个统一的管廊，然后由各个部门共享管道中的空间呢？就因为没有市政建设的经验，不知道一个城市最终要铺设多少管道，所有没有预先做规划。

由此可见，市政建设中的结果导向实际上是“头痛医头脚痛医脚”。建马路的时候，就应该预先建好地下管廊，这样才能避免后面不断地“开膛破肚”。

虽然现在大数据分析难以像公路管廊的建设，有一个明确的规划。具体的分析到底能产生哪些结果，现在还不好预测，但起码知道可实现的目标是多样的，因此需要留有扩充余地。所以在大数据建设时，不是先预设它的结果，然后再采集数据，而是要把可能采集的数据都采集出来。不能过于局限于采集当前所需要的数据，而是要把采集的范围放宽松一点，以便不时之需。虽然数据的采集和存储需要成本，但相对大数据的价值来说这种成本非常有限，而且成本会越来越低。数据的历史价值有的时候要大于实时数据的价值，所以在以后分析的时候，如果找不到历史数据，损失会更大。

由于当初不知道到底有多少管道需要铺设，如果提出建设地下管廊，大家就会质疑建立地下管廊的价值，让人怀疑是否过度建设。同样，作为大数据应用，前景非常广阔，完全应该接受类似教训，不要一定有明确的结果，因为即使现在能预测出几个结果，也并不等于是全部结果，很可能最后它的实际效益远远超过了当初的预计。

假如在以后某一天，大数据应用成熟了，需要哪些数据，每个数据有什么作用，能产生什么价值就会一清二楚，就像现代城市的地下管廊里用哪些管道基本上已经固定，很难再突如其来出现一个新的管道。但是，这种情况出现的时候，大数据应用已经成熟，如果你在这个时候才开始研究大数据应用，说明你在这个行业中已经落后了。

2.3.3 从有方向到无方向

在大数据应用早期，都是一些个性化的应用，这些应用为了满足某个特定的目标，称为有方向的数据挖掘。但个性化的应用成本比较高，应用面比较局限，所以希望寻找一个通用技术来应用数据，这就引出无方向数据挖掘的需求。

有方向的数据挖掘有非常大的局限性，因为数据挖掘的需求多种多样，而每个算法目标比较单一，不能适应这些多样化的需求。即使找到满足某种需求的算法，也有许多的参数需要调整，对使用者的专业化水平要求比较高。调整参数需要使用者有一定的业务知识，既要懂业务又要懂技术，这种人很少，所以影响到数据挖掘应用的发展。

如果有一个司机提供一种服务，即每天下午六点从上海浦东机场到陆家嘴，只需要 20 元，而一般打出租车可能需要 140 多元，这种服务有没有价值呢？

从表面来看是有价值的，但实际价值不大，因为它既设定固定线路，又设定了固定的时点，如果你想五点或者七点走，又或者不是到陆家嘴而是到外滩，那么就享受不了这种服务。由于恰好满足这个线路和时点的概率太小，所以一般人根本用不到这种服务。

这种情况非常类似于数据挖掘，或者准确地称为有方向的数据挖掘。因为有方向的数据挖掘是为了解决某一特定的问题，从而选择了特定的算法，并且设定了特定的参数，这种问题如果在工作中经常碰到、重复发生的，那肯定价值很大，但如果发生的不是这个问题，或者相关的参数发生变化，这个模型可能就不适用了。即使可以用同样的模型，但参数需要调整，鉴于调整参数需要经验，必须经过一些试验工作，原来开发的应用还是不能发挥作用。

鉴于在实际情况中需求是各种各样的，所以针对某一个特定的需求开发，它的价值总是有限的，就像这种特定时间、特定线路的汽车服务一样，限制因素越多，它的价值越小。

即使成功地开发了一个有方向数据挖掘的应用或产品，也只能适合一

个行业或一个企业的特定需求。在一个企业的众多需求中，也只满足其中一个。要满足多个行业、多个企业的多个需求，只能是一项不可完成的任务。

现在可以说，按照有方向数据挖掘的思路已经走入死胡同，因此许多著名软件公司放弃了这方面的产品的后续改进。

在发现有方向数据挖掘的缺陷以后，提出无方向数据挖掘的概念，也有人把它称为探索性数据挖掘。它的核心是不需要设定目标就可以进行数据挖掘。因为从事技术开发的人员不懂业务，即使懂业务的人也不知道自己挖掘的目标是什么，它需要根据业务的变化和当时的环境来确定挖掘的目的。

无方向数据挖掘相对于有方向数据挖掘，在开发的时候并没有一个固定的目标，它可以适应多种数据挖掘的需求。现在比较流行的自助 BI 工具，实际上也是实现类似的概念，但是因为它们以可视化为主，没有提供很多的算法，所以没有数据挖掘的应用层次深。但是，自助分析工具的走红跟无方向数据挖掘的需求是分不开的。

原来的 BI 展示都是由 IT 人员开发，IT 人员需要了解业务人员的需求，也就是说它要经过调研、开发、部署、培训整个一套流程，就像传统的软件开发一样，但是这种流程走下来以后发现业务部门的需求又变化了，最后导致原来的开发不能满足新需求而要重新开发，这个流程又需要走一遍，所以造成了 BI 项目的适应性不强。

为了解决这个问题，索性把 BI 的展示工作交给业务人员来做，由业务人员根据实际的需求，通过一个简单直观的工具来自己制作，类似于事务处理软件中的自定义报表。这样相对 IT 人员来说就更接近业务需求了，所以这种 BI 的自助分析工具也可算作一种无方向的数据挖掘。Excel 是最常用的自助分析工具，目前市场比较推崇的是 Tableau。

2.3.4 自助分析工具与自助分析系统的区别

在 BI 方面比较流行自助分析工具，主要是可视化的数据工具。通过拖拽操作，就可以从一个数据集中把数据设计成各种可视化的图形。

自助分析工具提供给业务部门使用，不像初期是由 IT 部门或外包公司

开发好以后再给业务部门使用，所以它称为自助分析工具。

自助分析工具虽然使用很直观，培训时间短，但它仍然需要业务人员学习使用，而且需要在本机上安装客户端程序，对如何从数据源中提取数据有一定的要求，可能还需要 IT 部门的人员去协助。如果数据源来自多个信息系统，或者数据源来自单个数据库的多个表格，或者数据需要经过预处理，这样工作业务部门的人员是无法搞定的。所以说它还只是一种工具，作为业务人员需要花一些时间来抽取数据和生成图形。

从理论上讲，业务部门工作重点应该是对已制作成统计图形的数据进行思考、分析，结合他的业务经验从数据中发现问题，并且通过对数据的钻取或者多条件的筛选来发现问题所在，从而及时地解决问题。这就需要提供一个自助分析系统。

自助分析系统是由 IT 部门开发好的，可以直接给业务部门人员提供可视化图形的一种系统，它不需要业务部门人员抽取数据和制作图形，比自助分析工具更为方便。

开发自助分析系统最基本的要求就是实现无方向的数据挖掘。在业务人员使用自助分析工具时，可以根据自己的需求选择数据和制作图形。作为自助分析系统，必须把业务人员可能需要的数据全部准备好，统计图形全部制作好，也就是说需要进行无需求的开发，提供过度设计才能满足业务人员的需要。

自助分析系统可以减轻业务人员数据分析的工作量，数据是由 IT 部门制作好的，可以很多的人同时使用，大大减少了重复工作，提高了工作效率。对业务人员使用自助分析工具的水平和熟练程度不再有要求，因此，自助分析系统更适合组织内部大量数据分析工作的需求。

可以直观地用餐厅的营业模式来说明自助分析工具和自助分析系统的区别。自助分析工具等于一组炊具，用这些工具的条件必须是自己设计或者选择菜品，采购并处理食材，拥有一定烹饪技术，才能制作出美味的菜肴。在推广自助分析工具时，销售人员常常会展示由专业人员利用这些炊具烹制的菜肴，让人看得垂涎欲滴而产生购买欲望，但它不是直接提供这些美味佳肴而只是卖“炊具”。

自助分析系统更像自助餐，客户根本不需要知道这道菜是如何烧出来的，用的是什么炊具和佐料，他的任务就是从这所有的菜肴中选择可口的去吃就可以了。如果觉得有些菜不想吃，可以不管它。这就是过度设计。

2.4 信息系统总体规划

2.4.1 基于数据的规划

一个组织的信息化总体规划非常重要，但现在大部分的规划由很多的独立子系统组成，子系统之间没有通过数据很紧密地关联在一起，就像人体只有分离躯干和肢体，没有神经信号交换。虽然有些总体规划把最高层注明为商业智能或者决策支持系统，但如何实现这些系统，还没有具体的方案，仅仅只是作为一个未来的目标。

在规划中首先要保证数据在整个系统中的畅通，明确不同子系统中数据之间的关系，而不是把不同系统的数据孤立起来，或者根本就不考虑数据。

虽然从技术上说，这些数据可以放到数据仓库里，但数据仓库只是一个存储的地方，目的是让数据能在组织内部共享，并且为组织的决策所利用，达到数据能够被最高层所认知并辅助决策的目的。

现在的信息化建设都是在做一些基础性的工作，主要是建设不同的子系统，用各个子系统采集数据，并且希望通过主数据的管理让数据统一，最终才会做到决策支持系统，即把决策支持系统当作信息化建设的最终目标。这种由下而上的信息化推进模式缺陷在于总体目标的缺失，并不能用这个规划来凝聚组织内各个部门的共识，无法团结大家来共同完成规划的目标。

一个好的规划应该有一个明确的战略目标，能够凝聚各方面的力量来共同完成。但是，如果规划只在纸上，只是信息部门的计划而不是一个组织整体的信息化计划，它涉及的只是 IT 部门的人，信息系统将难以体现其价值。如果一个组织的领导层和除 IT 部门以外的其他部门看不到与自己日

常工作的关系，信息系统的实施则难以得到其他部门的支持和配合。

作为一个组织的信息规划，应该让所有部门的人看得懂，起码要都能说得清楚这个信息系统实现以后会有怎样的效果、对目前的工作有什么样的改进、每一年工作完成以后会看到怎样的阶段性成果，即是要变成整个组织的信息化规划。

因此，整个组织的信息化规划应该从上而下，就是从公司高层对数据的需求出发，并且把数据需求分解到不同的职能部门，让他们认识到数据的重要性，然后会发现在现有的信息系统架构下这种需求是不能满足的。信息部门的工作就是弥补这样的缺陷，这样就把信息系统的建设和业务部门的日常工作联系起来。

比如说一个销售部门在日常工作中需要 100 个数据，按信息系统的现状可能只能提供 50 个数据，根据规划每年能增加 10 个数据，这样的话 IT 部门工作和业务部门就结合起来了。

总体规划的重点就是从公司高层决策支持者的需求出发，分解到部门的决策支持，然后落实到具体的信息子系统。也就是说，上什么样的子系统，并不是供应商有什么样的系统就上什么系统，或者其他公司上什么系统就上什么系统，而是需要根据业务部门对数据的需求来确定。规划不是以子系统为核心，而是以数据为核心，从数据的需求出发，到能够提供数据的信息子系统的开发、部署。

在总体规划中，一个制造型企业，数据可以分为四个层次：第一个层次是财务数据层次，核心是财务软件；第二个层次叫经营数据层级，核心是 ERP 软件；第三个层次是生产数据层次，核心是 MES 系统；第四个是控制数据层次，核心是工业控制、物联网。这个数据的汇总程度是从高到低，财务汇总的最高，控制数据层次最低，其他一些系统都是围绕数据的需求逐渐来弥补它们的不足。

规划不仅仅是在纸面上可以多个子系统组合在一起，相互之间有上下级的关系，有线条关联，最核心的是这些子系统之间的数据可以相通，数据可以逐级汇总。数据可以从上到下查询或分析，可以向下钻取，由粗到细。达到这样一个过程的规划才能称为一个完善的信息系统总体规划。

2.4.2 用规划展示数据不足

在做信息系统规划时，用数据作为一个联络不同子系统的神经信号，可以从总体上连接不同的信息系统，形成一个统一的整体。为了让规划被组织的其他部门和领导直观认识，可以仿造在城市规划中搭建模型的方法。

在城市规划中，会建一个三维的模型，然后把规划中哪里会建商场或写字楼、哪里会建居民小区等用三维模型显示出来，这样可以给参观者展示出未来城市建设的美好前景。

在信息系统规划中，可以把信息系统建完以后得到的各种数据指标展示出来，做成仪表板的形式。数据可以模拟。在仪表板中有些数值不能动，比如说一直是“0”，就像在规划模型中很多楼是假的一样，但有些已经可取得的数据就可以正确地显示处理。这样，就知道信息系统哪些是已经完成，哪些地方还有待完成，在以后需要上新系统的时候，就能够马上明白这个系统对完善仪表盘具体有哪些作用，可以为哪些指标提供数据，这些指标对经营管理是否重要。

大家知道既然是规划，那么肯定要超前。很多人认为自己单位做数据分析的时机不成熟是，原因是没有数据，这种说法放到城市规划中就是楼没有建好，所以无法规划，也就否定了规划超前的作用。规划的时候没有形象的展示，就像城市规划没有模型不容易被别人理解一样，会变成信息部门内部的一种自娱自乐。

在信息系统规划中，展示数据的不足不是一个缺点，而是一个优点。首先展示的整体目标能让人看到未来，展示现在的不足证明了信息系统的工作的重要性，证明信息部门还有很多的工作需要做。

2.4.3 以市长为核心的智慧城市总体规划

在智慧城市建设中，整体规划非常重要。许多城市在建设智慧城市的时候，也邀请了咨询公司来做总体规划，但这些总体规划没有以数据为纽带把子系统统一起来，没有把作为城市决策者——市长的需求作为整体规

划的核心。

智慧城市的数据并不是孤立的数据，而是相互关联的数据，要把数据当作各子系统之间连接的一个纽带。智慧城市需要建立一个数据仓库，把所有的数据都放进去，而数据仓库是统一的。数据仓库既可以作为基础数据库，作为多个子系统共享数据的来源，也是子系统运行的结果，把不同子系统运行后的结果汇总在一起，供市领导决策或其他部门共享。

建立数据仓库要把子系统建设中的程序和数据进行分离。程序以满足不同的职能部门的需求为主，但不管它有什么功能，生成的数据必须能够抽取到数据仓库中，要能被其他部门共享。

智慧城市中的不同子系统，虽然用互联网实现联通，就像人体一样，它的神经网络是通的，但里面缺少类似神经信号的数据来传播，所以它们之间并没有建立起一个相互关联的关系，缺乏一个协同的作用。

智慧城市的建设应该围绕市长决策需求，设计建立一个市长指挥室，以指挥室为智慧城市建设的抓手有两个好处：第一个好处是让市领导对智慧城市建设状态有一个直观的了解；第二个好处就是智慧城市的建设能边建设边发挥作用，并且随着智慧城市的深入推进这个作用会越来越大。

指挥室主要有哪些功能呢？指挥室要按照智慧城市完工的格局来布局，把整体架构先搭好，能显示一个城市方方面面的数据。这个架构中包括实时的指标显示，以及历史数据的检索。指挥室首先有一个控制中心，从每一个领域中提取一些重要的数据显示，比如说房产成交量和成交均价变化是地方政府非常关注的指标，也是制定调控政策的重要依据，因此市政府领导需要关注房产最新成交量和成交均价，及与去年同期的同比数据，数据应该每日更新，每天可以看到前一天结束时的数据。详细分析可以看到年初至今的累计成交量、成交均价，本月的成交量、成交均价，按时间维度分析每月的成交量、成交均价变化，按空间维度分析城市各区的成交量、成交均价。

这种设计在智慧城市建设初期，数据肯定是不全的，甚至可能只有小部分的数据。对于一些不能提交数据的职能部门，主管领导会产生压力，他就会努力督促下属完成信息化建设工作，并且把数据送到市长指挥室。

市长在平时看数据或开会的时候，就可以明显地看到哪些数据是新增的，对下属的工作起了一个反馈和激励作用。通过这种不断地反馈和督促就能大大推进智慧城市建设的步伐。市领导看到这些数据在他的日常工作中发挥了作用，他对这项工作也会越来越支持。由于对投入的产出心里有数，他支持的力度也会越来越大。

比如，在需要对房产市场进行调控的关键时刻，房产管理局及地方税务局不能提供及时的数据支持，市长不能看到成交量和成交均价的数据，并得到分析结果，显然他对两个局的工作会不满意，而政府的决策也难免会出现偏差。

目前的智慧城市建设主要从政绩观或者上级领导要求出发，主要目的不是利用智慧城市提升城市管理能力，而是通过智慧城市建设的需求来吸引一些相关的软件公司落户。由于政府没有切实的需求，导致这些落户的企业工作没有目标，最后它们的工作的价值也没有得到体现，这很不利于整个软件产业的良性发展。显然产业的良性发展就需要软件公司开发的软件能够在市政管理中产生实际作用，从而带动更多的需求。对于政府来说，通过提高效率、节约投资等来获取收益，反过来会督促软件企业开发出更多更好的产品，这样才能形成良性循环。

如果一个市领导在大小会议上宣传智慧城市的作用，而实际上他自己又没有在工作中获益；他一方面在开会的时候宣传信息化的重要性；另一方面他又没有用任何的信息化手段，那么他的下属部门负责人也不会特别重视信息化。目前智慧城市的建设是政府通过行政措施推动的，政府制定规划后由下属职能部门分别负责开发，由于开发的周期比较长，开发的结果又不能及时汇总，最后到市政府层级没有得到任何有效的反馈，这不利于整个任务的推进。这种情况跟企业信息化类似，老板强调信息化的重要性，并且要求下面搞信息化，自己也投钱，最后实现信息化的人都是下属而非老板。智慧城市最坏的结果就是一般公务员都在信息化，而市领导没有。这个问题出现主要还是决策支持系统的开发比较落后，作为领导如何利用决策支持系统去工作还没有形成一个标准的模式。





第 3 章

推动数据革命

全球范围内，运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势，有关发达国家相继制定实施大数据战略性文件，大力推动大数据发展和应用。

国务院《促进大数据发展行动纲要》

3.1 数据的立法

要迎来数据时代，除了技术上工作以外，还需要立法的配合。

第一是数据安全性的立法。

立法的前提是明确到底要惩罚什么？是数据的公开、数据的传播、数据的应用，还是利用数据产生的后果。

如果把立法重点放在数据的公开、传播和应用上，由于无法把握合适的尺度及同一个行为可能产生两个以上的后果，导致对合法行为的误伤，影响数据的利用。

所以，立法应该把重点放在数据不当应用产生后果的惩罚上。好比不能立法禁止菜刀的生产、买卖和使用，但如果有人用菜刀杀人，那肯定入刑。

第二是立法的重点在于数据强制性公开。

如果一个企业享受了政府给予的垄断权力，它就有义务把财务数据和经营数据公开，这与公司上市需要定期发布财务报告的道理一样。立法虽然不能强制所有的企业和个人把数据公开，但它只要有任何权利要求，都必须公开相关经营数据。

虽然证券市场的建立是在信息公开的基础上，但它每季度提供的定期报告信息过于粗略，无法发现问题。美国的《萨班斯法》就是在“安然事件”发生以后，为了加强了对上市公司信息披露的要求而制定的。但是，这种要求增加了上市公司的负担，而且这种公开也不彻底，所以并没有真正解决上市公司信息公开的问题。

经营数据的公开可能会影响到企业的商业秘密，但是，一切用信息不对称建立的核心竞争优势都是不合理的。数据公开立法的目的，就是消除所有通过信息不对称来获取利润的途径，从而使社会的资源得到最为合理

的配置。从海关数据公开的多年经验来看，进出口数据的公开并未影响相关企业的竞争力。

利用信息不对称来获取利润的现象，在很多行业都存在。比如银行的信贷业务就是利用了借款人和贷款人信息不对称，现在的 P2P 平台就是要消除这种信息不对称。

第三是数据作为起诉证据的时效。

如果一个被告将自己拥有的数据公开，并且数据的产生时间在一定时间（比如两年）前，则任何人都不能以这些数据作为证据控告他。比如，一个公司公开了两年前单位工资数据，有个离职员工根据这些数据，发现公司没有严格履行《劳动法》，没有为自己足额缴纳养老金，并据此向法院起诉要求赔偿，法院可以拒绝此项证据的有效性。如果企业没有公开数据，或原告的数据不是来源于开放数据，则不适用此项法律。

此项立法有利于鼓励组织公开数据，从而推动社会的开放数据建设。

3.2　数据的公开

3.2.1　对信息公开的认识

数据安全已成为数据公开和应用的最大的障碍。数据安全要考虑，但按照目前数据的开放和利用的程度来说，考虑为时尚早。

可以比照食品安全问题去看数据安全。现在大家都比较重视食品的安全，设想一下，当遇到灾荒的时候，即食物供应严重不足的时候，大家还会考虑食品安全的问题吗？肯定是在食品比较富裕或者过剩的情况下才会有这个问题提出来。

如果把数据比作食品，现在处于饥荒中，可用的数据很少。虽然确实出现一些数据泄露以及恶意利用数据的问题，但并不能阻挡信息公开，就像不能因为发生交通事故就禁止汽车一样。

利用菜刀杀人的情况是永远存在的，但是，不能因此就在买卖菜刀的时候先考虑菜刀杀人的问题，而应先考虑菜刀可以切菜切肉等生活问题。

对信息公开来说，这个数据安全问题完全是过虑了。

影响数据公开的另一原因是数据公开的价值还没有被人们认识到，否则数据公开就不会遭到如此反对。假设有一个癌症病人，需要和他签订一个协议，将他的癌症详细信息公开，虽然现在公开并不能直接挽救他的生命，但对后面的病人，医生可以利用他和其他病人的病理数据进行研究，从而找到治愈的方法。也就是说，如果我们的数据能够早点公开，医生能够提前五年得到数据的话，可能就找到方案给该病人治疗了。如果让癌症病人来选择，也许很多病人都会来签订这个协议，同意对数据进行公开。

同样的道理，超市想和你签订一个协议：假设你同意公开你的购买信息，可以降低商品 5% 的价格，也许很多消费者也愿意签订这样的协议。实际上现在的网上购物，就已经把自己的信息在网上完全公开。为什么很多人愿意，其中一个原因肯定是由于网上的价格比较便宜。

所以说，数据公开不是一个安全问题，而是一个利益问题。我们只要把数据公开跟信息拥有者沟通清楚，他们并不会反对信息公开。如果现在能制定相关的法律来限制对信息的恶意利用，更能消除大家对信息公开后产生的疑虑。

3.2.2 政府开放数据

自 2009 年美国总统奥巴马签署了《开放透明政府备忘录》和数据门户网站 Data.Gov 上线以来，数据公开的必要性已经在西方国家达成了共识，世界各国逐渐走上了开放数据的道路。虽然开放数据已经体现按原始数据公布和提供可机读数据这两个重要的开放特点，但是还是遇到很多难题。第一个难题是政府对这些数据的利用没有一个明确的方向；第二个难题是数据粒度还不够细化，影响了数据的使用价值。

由于政府并未对数据的用途、使用方法、处理技术和所能产生的价值有一个明确的意见，所以只是通过开放数据把这些问题提交给社会探索，

并将利用数据开发的软件在政府开放数据网站进行推荐。数据虽然得到开放，但缺少数据应用的成熟思路。

政府的数据只是统计数据而非原始数据，比如学校的数据是按学校统计汇总的数据，而不是基于学生个体的详细数据，个体的数据存在于学校而不是政府。数据汇总后，失去了许多维度信息，会影响数据的利用价值。

我们知道，数据的价值是通过支持决策产生的。如果没有决策行为或者决策中没有用到数据，那么这些数据是没有产生效益的，或者决策中看到这些数据，但数据未能准确地传递含义，那么数据也是没有产生效益的。所以，仅靠现在的政府公开数据，我们还无法发现和实现其价值。

在开发一个信息系统时，如果涉及大量的数据，会有很多成本花在数据整理上。现在政府开发数据一般按时间用文本格式（如 CSV 格式）提供，文件数量比较大，软件开发公司需花费时间对数据做处理，这影响了开发进度和增加了开发成本。

因此，政府开放数据的下一个目标，首先，就是要把分布在不同文件中的数据整合到一个分布式数据库中，通过一个入口进入。其次，数据可以通过一个比较常用的语言（如 SQL），或者提供一个 REST 接口查询。最后，是细化数据，比如学校数据要细化到每个学生。明细数据可以不存在于政府，可以存在各个学校里面，或者学校通过委托运营商放在云服务器中，由政府统一提供访问接口。

3.2.3 对开放数据的要求

作为开放数据，应该达到什么程度才能真正能发挥作用？

举例来说，一个城市里有一个图书馆，目前一些地方政府的开放数据网站虽然提供了图书馆静态信息，比如地址、联系电话、开闭馆时间，但这些种信息不具备开发数据所需的价值。

信息价值最基本的特征是时效性。简而言之，就是数据记录里面必须有一个信息是跟时间有关的。那么，图书馆里跟时间有关的信息是什么样的呢？首先可能会想到图书馆一天固定的开馆和闭馆的时间，或者图书馆

闭馆休息的日期。但这些数据不是一个动态的记录，没有特定日期开闭馆记录，不具备时效性，只有图书馆工作人员对每天开馆和闭馆时间的记录是有时效性的。

图书馆最重要的开放数据是图书的借还记录，包括借书的时间、借书人的信息、书籍的信息、还书的时间。图书借换记录表记录了主要数据，称为事实表，另外还需要两张相关表，称为维度表。其中一张维度表是关于借书人的信息，包括姓名、出生年月、性别、受教育程度等，当然为了保密起见可以将姓名隐藏。另一个维度就是书籍的信息，包括书名、出版社、图书分类等。

这种数据如果公开，对研究该地区人们阅读习惯和兴趣点有很大的作用。利用这些数据可以分析这些地区人们读书爱好的类型，读书的目的，当前社会上阅读的热点等。证券公司如果发现一个地区阅读股票类书籍的人比较多，证明该地区市民对股票比较感兴趣，就可以多布局证券营业网点及加大推广力度。

比较不同城市的阅读倾向，可以明显区别不同城市人们的偏好和人文情况。现在有人研究中国的城市群得出一些结论：长三角是中国经济发达、文化修养比较高的地方，珠三角经济虽然比较发达但文化层次偏低。这些依据从哪里来？如果有这些阅读的数据，显然就比较好研究了。

3.2.4　政府主导的公共数据库

2016 年 1 月，美国总统奥巴马宣布发起一项寻找癌症疗法的大型计划，用“登月计划”作比来说明其重要性，后来又设立以副总统拜登为首的“白宫抗癌登月计划特别小组”，目标是让抗癌的研究进展速度翻一番，在 5 年内取得原本可能要 10 年取得的成果。

美国副总统拜登 2016 年 6 月 6 日宣布，启用癌症“登月计划”首个大型开放数据库，旨在更好地分享癌症相关数据，帮助全世界科研人员更好地认识癌症，从而开发出更有效的疗法。美国白宫当天发表一份声明说，这个名为“基因组数据共用”的数据库将为储存、分析和分享癌症基因组

数据及相关临床数据提供一个公共平台，这对推进精准医学、改善癌症治疗水平至关重要。

白宫声明中提到，这个数据库是一个交互式系统，提供的都是未处理过的原始数据，科研人员可以使用新研发出的计算工具与分析方法重新分析这些数据。数据库整合了美国国家癌症研究所现有多个癌症研究项目的资料，目前已拥有 1.2 万名癌症患者的数据，今后全世界科研人员可继续向其中添加更多数据。

“基因组数据共用”数据库将由芝加哥大学代美国国家癌症研究所管理。该数据库已经具备数据时代的一些特点：

（1）数据库属于国家癌症研究所，它是一个公共事业，不是私营企业行为；

（2）数据库由芝加哥大学代为管理，而芝加哥大学是一所私立大学，即要利用民间组织的力量；

（3）数据库中是原始数据，而不是经过处理的数据，保证研究的灵活性；

（4）数据库是开放的，全世界科研人员都可以添加数据；

（5）相对攻克癌症挽救生命来说，患者的隐私已不再重要。

3.2.5　科研数据的公开

科学数据（或研究数据）是指在科技活动中（实验、观测、探测、调查等）或通过其他方式所获取的反映客观世界的本质、特征、变化规律等的原始基本数据以及根据不同科技活动需要，进行系统加工整理的各类数据集。英国数字保存中心（digital curation centre，DCC）认为数据管理与共享具有多重益处：

①在需要使用数据时，用户能够找到并理解数据；

②当有研究人员离开团队，或有新研究人员加入时，能够保持工作的延续性；

③用户可以避免不必要的重复工作，如重新采集数据；

④支持文献的数据得以保存，从而可对文献结论进行验证；

⑤通过数据共享可以开展更多的合作，推动科学研究；

⑥能够提高研究的显示度；

⑦其他科研人员可以引用数据，使数据拥有者获得更多荣誉[2]。

在开放获取（open access）理念指导下，科研资助机构积极推动其资助的科研产出的开放获取。以往科研资助机构主要关注期刊论文、会议论文等正规出版物的公开获取，近年来以数据为中心、数据驱动科研的特征越来越突出，为保证科学研究的完整性，科研资助机构开始促进作为科研产出组成部分的研究数据的共享与开放获取，并制定数据管理与共享政策。科学数据管理与共享政策的制定是科学数据共享工作顺利进行的保障，也是推动科学数据管理与共享的主要驱动力之一。在科研资助机构的数据管理与共享政策的要求下，研究型图书馆及大学图书馆开始为研究人员制定数据管理与共享计划提供支持与服务。[5]

3.3 有时数据隐私只是借口

数据的应用现在刚刚开始，它的前景如何、价值如何，社会上还没有形成一个共识。

目前大家对于数据的只是隐约地认为它有很大的价值，至于具体它对社会经济有什么作用、如何发挥作用，社会或者个人愿意付出多少代价实现这个目标，这些问题都是未知。

如果现在把主要精力集中在数据的安全性方面，就像先建起一堵挡在通向数据未来的墙，大家都在研究如何推倒这堵墙，而无暇关注这堵墙后面有什么。

任何人类的探险活动都有一个伟大的目标在激励着。哥伦布穿过大西洋发现新大陆时，没有人知道他会遇到多少风险，但他有一个伟大的目标就是前往东方，获得东方的香料。因为东方的香料在西方具有很大的价值，所以他得到西班牙女王的支持，获得资金；说服船员加入他的队伍进行未知前程的探险，所以他成功的关键是有伟大的目标。

在数据应用的探索中，我们的目标现在还不太明确，就被数据安全这堵墙挡在路上，也许以后想起来会觉得很好笑，因为我们失去的隐私与所得到的相比可能是微不足道的。

现在，我们的主要任务应该是描述数据应用的未来，研究通过何种途径能够产生最大的价值，给大家描绘一个虽然失去部分隐私但能拥有更多的未来。

当然，隐私需要保护，只是相对于我们的目标而言，它已经不很重要了。比如食品安全，没有人可以说它不重要，不安全的食品会造成生病、中毒乃至死亡这样的后果，但在食物贫乏的地区和年代，饥饿的人们会忽略食品安全问题，因为身体的健康与死亡相比是微不足道的，所以食品安全会退居次要层面。

同样，在数据安全问题上，假设一个人得了癌症，他的个人信息和疾病症状、用药信息，本人 DNA 的信息都是个人隐私，但如果有人承诺，只要同意公开、出售这些信息就能够治好他的病，那么估计没有病人会拒绝。

中国的一家医药连锁集团代理一家国际医疗公司的产品，这个医药集团在国内约有 50 亿元的销售额，它在国内的销售渠道和销售对象应该都是很有价值的数据，若是被国外厂商知道，会降低它的价值，但国外的厂商又想通过这些数据研究产品去向，对产品下一步研发起到作用。开始时，经销商不愿意给出数据，这是典型的数据安全问题。然后国外厂家提出条件，若是经销商提供了这些数据，它会返 5% 的销售额。在这种利益诱惑下，经销商丢掉数据隐私担忧，积极地开发数据信息系统为国外厂家提供数据。这个案例中表面是数据安全问题，实际上是该集团手上并没有现成的数据，而采集这些数据需要花费不少成本，在国外厂家解决了成本问题后，数据隐私的障碍也就不存在了。

所以数据是有价值的，当价值得不到体现的时候，常常被人以隐私作为借口。但当数据有交换价值的时候，很多隐私问题就不再存在。

在医疗改革中，获取病人的治疗信息非常重要，不论是跟病人还是和医院洽谈病人的治疗信息，他们都会以隐私为借口拒绝。对于病人，如果能够给他们一定的费用来交换，或者从社会角度给配合交换隐私的人减免

5% 的医药费，那么很多人会乐于提供信息。

现在很多数据提供要求被拒绝，关键在于提供者看不到回报，提供数据之后没有得到直接的利益。数据应用的关键是让数据相关者了解数据应用后产生的价值，并通过数据产生的价值回报数据的提供者，从而形成良性循环。如此，数据产业就会得到理想的发展，而在确切的回报没有得到之前，能够描述回报的前景更为重要，这样才会集聚更多的社会资源投入到开发中。

事实上，更多的时候数据隐私是一种借口。按照现在的信息技术水平，提供一个比较准确的数据需要做很多的工作，包括数据相关软件的开发、业务部门的数据准确录入及数据最终的核对和定时提交，无论是额外支出的费用还是内部人员的配合都需要巨大的成本。即使已经有了数据，但这些数据与其他数据混合在一起，对数据的分离和单独传送也会产生一笔额外成本。开发一套软件或者招聘专人去核对和提交数据是较易估算的显性成本，但业务人员配合工作所花费的时间是隐性成本，一般难以明确估计，所以作为数据索取者应该充分理解数据提供方所花费的巨额成本。倘若数据提供方觉得无利可图，是不会投入成本的。这时，他会以隐私作为借口。

有一些组织，信息化水平比较落后，无法获取数据，但又不能承认他们没有能力和数据，他更会以数据安全为借口予以拒绝。所以针对数据安全这一问题，应该找到安全后面的真正理由，从而有针对性地解决。

3.4 数据基础设施

推动建立政府部门和事业单位等公共机构数据资源清单，按照“增量先行”的方式，加强对政府部门数据的国家统筹管理，加快建设国家政府数据统一开放平台。制订公共机构数据开放计划，落实数据开放和维护责任，推进公共机构数据资源统一汇聚和集中向社会开放，提升政府数据开放共享标准化程度，优先推动信用、交通、医疗、卫生、就业、社保、地理、

文化、教育、科技、资源、农业、环境、安监、金融、质量、统计、气象、海洋、企业登记监管等民生保障服务相关领域的政府数据集向社会开放。建立政府和社会互动的大数据采集形成机制，制定政府数据共享开放目录。

国务院《促进大数据发展行动纲要》

3.4.1　数据作为基础设施

数据的采集、保存和服务有一定成本，而数据服务有一定的公共性，因此数据建设应作为一种基础设施。公路、港口、铁路是属于工业时代的基础设施，网络和互联网是属于信息时代的基础设施，那么公共数据库则可称为数据时代的基础设施。

基础数据的采集，应该成为政府的一种公共投资。数据采集工作可以委托给私营机构来操作，但数据开放应该是一种强制行为，可以通过国家定价的方式来确定收费价格，从而给予数据运营企业一定的回报。

2016 年中国发生一起关于药品电子监管码（以下简称药监码）叫停事件。国家食品药品监督管理局要求所有企业推广药品电子监管码的决策受到了众多药品零售企业的反对，究其根源是这些药品零售企业不明白数据采集的成本应由谁承担，最终的数据如何为公共事业服务。政府通过行政命令要求企业来承担这方面的成本。山东某大药房连锁股份有限公司的董事长认为若要全面实现药监码其 760 多家门店合计一次性需投入 1 000 多万元，可是投入之后数据都为阿里健康公司所有，并且阿里健康会对这些数据进行二次开发和销售。显而易见，这会产生利益不对称的问题，药品销售连锁企业出资提供数据而可能被阿里健康拿去销售。

合理的方案是，首先，应规定数据为国家公共所有，仅由阿里健康负责运营，同时阿里健康对数据的使用定价必须公开，而不能以隐私为名行自己的盈利之道；其次，应用数据的定价应该实现政府定价，数据使用企业必须承担一部分费用。

阿里健康的数据变现，市场有一个培育过程，短期内的收入不可能完全覆盖整个数据的采集、储存和开发的成本。这种情况下，需要国家把它

当作一个基础设施来投资，而不能要求所有的投入都得到短期的回报，最低限度下，应该按照30年或者50年的运营期来计算投资回报。

投资回报除市场培育问题以外，数据的权威性、开放度不够也有很大的影响。阿里健康的数据提供的是单个数据检索的结果，或是经处理后的统计数据，不提供未经处理的原始数据。它按照自己的理解和开发水平来提供数据，而并非抱着开放的态度，或不是培育数据产业的目的同社会进行密切合作共同开发数据。由于阿里健康合作门槛高，一般企业难以与它合作，因此对数据的垄断导致了创新不足，同时又不给社会创新的机会，这就影响到整个数据产业的健康发展。

3.4.2 数据垄断的“滑铁卢”

2016年的药品电子监管码事件，是阿里健康试图垄断数据资源，遭到药品流通行业抵制后遇到的“滑铁卢”事件。[6]

药监码平台建设是数据社会数据基础设施建设的先行者，暴露出来的问题也是阻碍数据共享平台建设的共性问题，解决这些问题在国家层面必须先有一个明确的战略。下面分析一下出现这些问题的原因。

（1）政府监管部门的不重视。药品监管平台在数据时代是非常重要的基础设施，但政府部门从开始到现在，并没有认识到它的重要性，也没有形成建设的思路和总体的框架，完全放手让社会公司去做，自己只有名义上的所有权，既没有专门的部门和人去运作平台，也没有对这个平台后续的使用和权益做明确的界定，基本是采用一种“放羊”的方式。

（2）平台的建设由社会公司自筹资金完成，政府对它也没有系统的补贴，更没有采取采购服务的方式资助，社会公司就认为这个平台是自己建的，所有的数据也应该都是自己的。

（3）作为平台的建设公司阿里健康过于贪心，想利用政府赋予的垄断权获取最大的利益，而且这个利益的获取是全方位的，包括数据采集过程中设备的采购、数据传送的接口、数据的二次开发和使用，所有利益都想通吃，而且是一种排他性的通吃，完全没有共享的理念，把数据资源当

作自己获取利益的一个手段，而不是把数据资源作为一个公益的资源，然后自己在增值服务上获利。

如果我们重新复盘这个事情，正确的应该怎样做呢？

（1）政府主管部门（国家食药监总局）应该成立或授权一个专门的机构，主管药品监管平台的建设。该部门主要起到一个监管、协调的作用，主要的目标是通过制定规则，协调各方面的利益，来保证数据的采集以及分享，保证对社会产生最大的价值。

（2）在政府机构下面应该成立一个专家委员会，对数据采集的格式、分享的接口提出自己的技术标准，平台以后的运作都要遵循这样的标准，标准应该向社会公开。

（3）平台采取社会运作的方式，面向社会招标，让有能力、有意愿的公司运作这个平台。一个公司运作有一个周期，比如 3 年或 5 年，到期以后要重新招标。从技术上，专家委员会应该保证这个平台即使换一家企业运作也能平稳过渡，关键技术不会被前一个运营公司垄断。

（4）平台的运作应该完全由政府出资，除了企业自己用的设备以外，只要跟这个平台有关的专用设备都应该由政府出钱，更不应该向企业收取平台的使用费、接口费等。

（5）数据应该公开，最好是完全向社会公开。如果有顾虑的话，也可以向审核合格的数据服务商公开，完全可以参照海关数据、证券交易数据的公开方法。

（6）要鼓励社会上的软件公司围绕这个数据资源进行二次开发，为社会提供更多的服务，而不要像有些公司一样自己垄断一个资源，既不向社会开放，自己开发的东西又满足不了市场的需求。

3.4.3　公共数据服务与中介

我们从房产中介的工作流程来分析一下数据革命以后的变化。

我们知道，房产中介是一种典型的中介服务，需要依赖大量的数据。这些数据包括房产的供应数据和需求数据，哪些人需要买房或租房，哪些

人需要卖房或出租，这种数据是中介业务开展的基础。简单来说，房产价格上涨的时候掌握房源数据很重要，价格下跌的时候掌握需求数据很重要。毋庸置疑，数据可以说是房产中介的核心竞争力，房产中介投入了很大的人力、物力。

现在很多人直接在网上发布房地产供需信息，但这些房源或需求信息在网上公开后，打电话来的不是最终用户而是中介。虽然信息发布者希望供需直接成交，省掉中介费，但实际上大量房屋买卖或租赁还是在中介撮合下成交的。二手车交易也有类似现象。

因此，我们可以得出以下两点结论。

（1）房产需求和供应信息的获取是很低效的工作，这个需要第三方或者是政府提供带有公共事业的角色去服务。

（2）在供需信息公开的基础上，并没有降低房中介的价值。

在数据时代，应该把供需的信息平台作为一个公共事业，但这个公共事业的发展并不会降低中介的作用。房产交易的撮合并不是有了供需信息对接就可以轻易完成的，还有很多工作需要做，中介可以提供很多相关的服务，比如提供保证金的担保、充当签约的一个中介、协助办理贷款和进行房地产登记。

当然，通过公用信息的提供来降低房产中介的收费，也是减少社会成本的方法。以后，中介的工作要建立在信息共享的基础上。其实这方面在中国的旅游业做得比较好。旅游业有组团社和接单社的区别，组团社负责组团，组团以后会把信息发布在网上，然后所有的旅行社都可以接单，最后都送到一起来，这样就能保证任何一个组团社能得到更大的客源，资源因此得到了整合。这种整合实际上是市场自由竞争的结果。

所以，在数据时代，应该更多地从数据的公共性着手去建立这个平台，从而降低整个社会的成本。

3.4.4　农产品交易数据的案例

我们以农产品交易国内市场的建设为例，说明如何利用信息技术和数

据分析来实现全国农产品市场一体化的建设。

数据一体化市场建设的核心是，建立农产品交易的公共平台以及交易数据的共享。建设主要包括三个原则：一是以社会运营为主；二是数据公开；三是公开提供数据分析结果。

首先，这个平台应该由国家成立专门的机构建立，并且国家拥有平台的所有权和数据的所有权，可以将其定义为公共事业。国家建立专家委员会制定和审查数据交换的标准。平台的运作应该由社会上的公司来运作，采用政府购买服务的方式，通过竞标来委托社会公司开发运营。平台的运营权可以转让，防止运营企业垄断。

其次，数据必须共享。数据虽然是这个平台运作的，但它的所有权归国家，只要任何单位和个人符合一定的资质要求，都可以获取数据，而且必须是最细粒度的原始数据，平台的运营商不得拒绝。

最后，对这些数据分析的结果应该公开传播，特别是通过新闻媒体来传播，保证所有的人都可以免费获得数据以及免费获得相关的分析结果。

对于数据公开，可以借鉴股票数据流转的方式。我们知道，全世界股票交易所的数据都是公开，有很多的公司能够提供股票数据服务，可以通过时间差来收费，也就是说，如果你要实时接收数据肯定要收费，如果你不需要实时，只要延期数据，可以免费。

数据共享，不能仅提供简单的数据检索方式，即根据检索条件出来几笔明细数据记录，这种数据价值不大，因为这种细粒度的数据，无法让使用者从整体上把握市场情况。

一个农产品经纪人会怎样利用数据呢？他会根据自己熟悉的品种、区域，在地图上发现同一种农产品价格的差异，从中赚取差价。例如，甘肃省某种农产品的交易价格为 10 元钱，同种农产品在广州交易价格可能是 20 元。这个数据不是从单笔数据上看出来的，而是从成交价格的平均值和交易的历史趋势发现的。在发现这个规律存在后，经纪人会计算中间的仓储和运输的成本，发现有利润，就会在甘肃采购产品，再把产品运到广州出售，实现盈利。

如果过多的经纪人将甘肃的这种农产品运输到广州来卖，结果不是导

致甘肃农产品收购价格的上升，就是导致广州农产品销售价格的下降，从而使中间的价差正好满足运输仓储的费用和适当利润的范围，达到一种平衡。

如果有这样的平台，农产品经纪人就能及时获取相应的数据并对数据研判，全国的农产品的价格就会很快趋于一致，不会出现由于信息流通的不畅导致库存的积压和浪费。

为保证数据的及时录入和准确性，可以采用会员制的方式，收集数据和录入的数据是同一人，约束数据造假行为。另外，可以把对农产品的补贴和数据挂钩。为确保农产品数据的准确性，可以考虑卖方数据和买方数据的同时录入，然后进行数据的互相校验。

通过这些数据，可以了解交通运输费用对农产品交易的影响，从而采取相应的对策。另外对农产品的补贴也可以有一个明确的依据和数据化的标准。

3.5 建立数据图书馆

我们知道的传统图书馆，主要以保存书籍为主，内容以文字为主，中间有插图。现代图书馆也珍藏影像等多媒体资料。

在数据时代，会出现与传统图书馆相似的数据图书馆。数据图书馆准确地说应该叫数据馆，因为只保存数据而不保存图书，但为了和现在的图书馆有很好的联系，就叫数据图书馆。

数据图书馆和数字图书馆有很大的区别。数字图书馆本质上和现在的图书馆内容上是一致的，只不过是把它数字化了，比如说原来图书是纸质的，现在电子化了。

那么数据图书馆呢？首先它储存的内容是数据，即阿拉伯数字以及相关的一些说明。数据图书馆存储的范围远远超过传统图书馆，而且它的数据是原始数据，研究价值更大。数据图书馆的原始数据与根据这些数据撰

写的书籍不同，原始数据更为精确。同一个数据，不同的人可以得出不同的结论。数据图书馆更像一个数据素材馆。

数据图书馆保存的内容，是在不同介质中的数据。数据图书馆第一个功能是读取数据，可以提供多种手段读取保存在不同介质中的数据，即使很老的一个软盘，也能读取里面的数据。

为此，除了启用一些老的计算机，进行维护后使用外，还可以开发新的技术，制造可以读取很多格式介质的设备。就像现在读卡器一样，同一个插口可以读很多类型的存储卡。

数据图书馆第二个功能就是对数据进行整理。数据图书馆的数据有两种利用方法，一种是对存储在原始介质中的数据直接解读，还有一种就是分布式的解读。

直接解读原始介质中的数据比较困难，除了读写设备外，其数据存储格式，元数据的内容常常未知。而且，阅读者必须在现场，无法远程获取数据。

应该把所有的数据整合到一起去。所以我们要提供一个分布式数据存储。提出分布式数据存储主要原因是数据不应该分散存在不同的介质上，特别是原始介质上，因为它的容量很小，而且容易出错。我们应该存在一个云存储上，这个云存储应该是分布在世界各地的，不一定局限于本地。这种格式是标准的，可以解读的，可以通过标准的 SQL 语言检索到分布存储在不同云中的数据。

关于数据图书馆中数据的来源，可以通过各自的捐赠或者购买来实现。大家觉得数据的安全性非常重要，数据拥有者不会愿意提供数据，实际上这里面也分多种情况。比如说有些数据原来的所属公司已经不存在了，它的数据库应该就可以公开了。还在正常经营的企业，它的三年或五年前的数据库也可以提供。

通过提供数据的备份托管服务，可以和数据提供者协商一个保管和开放的时间期限。比如说有些企业信息系统升级了，老系统数据是放在老服务器中，过几年以后，服务器中的数据可能读不出来了。即使数据可以读出来，随着人员的流动，过了几年老员工离职后，这个数据也无人可识别。

如果该企业把数据捐赠到数据图书馆，就有专业的技术人员给它保护、维护、解读。当公司几年以后仍然需要这个数据，还可以得到，显然对公司有利。有些数据做一些经过公司认可的处理，去掉客户的名称或者员工的名字等敏感性信息，用编码代替，那么这个数据就可以立即公开。没有去掉的可以在五年至十年后公开。通过这种和企业的协议，解决数据的隐私问题。

关于数据的共享，可以采用对等开放的原则。对于共建分布式存储的协作组织内部，数据对等互相开放。其他个人和单位通过协作成员访问，通过协作成员对访问进行管理和控制。

数据图书馆中数据使用的一个重要规则，就是数据不允许全集拷贝。可以下载其中一部分数据或汇总数据，但不可以原样复制。设置这个规则的理由是：

（1）数据已经公开，其他人需要这个数据发链接过去就可以了，根本不需要下载或复制；

（2）数据有一致性的问题，复制后数据有可能被篡改；

（3）保证资料的唯一性，保护数据收集者的权利。

下面的一个故事描绘了数据图书馆使用的场景：

王素是中部地区一所著名综合性大学经济系的研究生，他正在研究的课题涉及员工收入在企业成本中的占比，需要统计企业有关经营和工资的微观数据。在2030年，高校论文没有以前这么简单了，经济学的论文要像生物医药的论文一样，必须要有详细的数据支撑，仅仅是概述性的东西导师是不会认可的，也没有地方愿意发表。

王素为了得到数据，来到位于长三角的一个小镇，它是中国数据图书馆的发起者之一，也是现在数据存储的主要云数据中心之一。这个小镇是由一个具有百年历史的工业基地改造的，很多机房和办公室都是由原来的老厂房改造而来的。

那么，王素为什么要跑这么远来到这个小镇上呢？数据图书馆虽然是一个基于网络的分布式图书馆，王素所在的高校也是数据图书馆联盟的成

员之一，它也拥有大部分数据查询的权限，王素可以看到分布在全国多地云数据中心的数据。但有些数据是一些企业的近期数据，还没有到授权对外开放的时间，这些企业的数据指定寄存在一个或多个数据图书馆中，若需访问只能在本地，而不能在网上共享。只有等到与这些数据提供者签订的协议中规定的时间，才能开放，这个规定时间一般是 3 ～ 5 年。因为企业为了避免过早开放对现在的经营或者与相关单位产生纠纷，所以设定这样的期限。国家的数据安全法也规定，根据数据诉讼的年限是两年，在两年以后不能把这些数据作为诉讼的依据，即使是作为诉讼的来源也不行。所以企业数据一般要在两年以后才能够上网。

王素为了在研究中得到最新的数据，他必须到数据寄存最多的数据图书馆所在地查询。他是如何知道这个小镇数据最多的呢？因为在学校的数据图书馆里可以查询到所有的数据目录和数据更新的最后时间，他在学校做了一些功课，将数据的目录列出来，并且对数据存储的地点进行筛选，从而发现这个小镇作为数据图书馆的发起人之一数据最多，所以王素决定到这个小镇来寻找数据。当然，如果他通过对搜集数据分析，发现他的论据还不够充足，也可以到其他数据图书馆去查询。

王素在这个小镇住下之后，第二天就凭着在学校里的数据图书馆证顺利进入数据图书馆。他用馆里的计算机，访问未接入互联网的私有云数据服务器，找到相关的数据目录，在数据目录中找到产生数据的软件供应商名称、产品名称、版本号，还有详细描述数据库结构的数据字典。

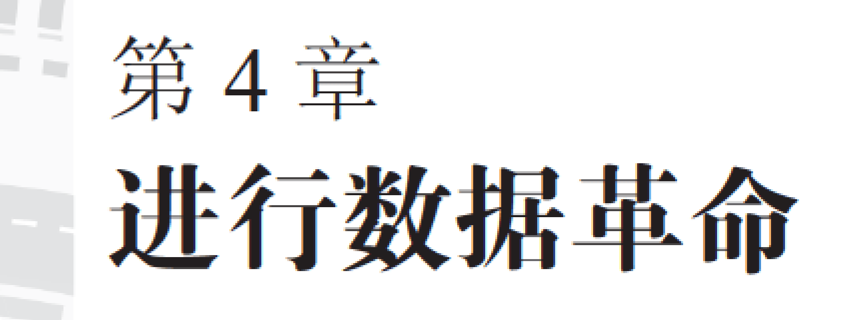
这些数据主要来自企业的 ERP 系统和工资管理系统，从 ERP 系统里可以看到这些企业每年的销售收入和成本支出，从工资管理系统里他可以看到员工人数、工资总额、人均工资及加班、补贴情况。他从中选取了十家企业，其中有五家用的是同一家软件公司的产品，还有五家分别用了两家软件公司的产品，即总共有三种软件。

王素根据其中一家软件公司产品的数据字典，编写了一些 SQL 语句来读取数据，经测试通过后，根据其他两家软件公司的数据字典进行修改，最后变成三组 SQL 语句。这样，他就可以执行这些语句查到相关的合计数据。

根据数据图书馆的规定，原始数据只能读而不能下载，但通过 SQL 语

向处理的合计数据可以下载，之后再通过专用的邮件系统发到自己的邮箱里。这个邮件系统是数据图书馆专门开发的，它有数据审核功能，主要是为了审核数据量，如果数据量太大就证明了你在下载数据，这是不允许的；第二个是对数据进行脱敏，有些涉及人名或者是企业名称的数据是不允许发送出去的。经过这样的方式，王素把取出的统计数据发送到自己的邮箱里，然后他可以通过可视化工具对数据进行分析或者生成统计图形，再把这些数据嵌到自己的论文里发表。在论文里他必须注明数据来自哪家数据图书馆，以及数据库的编码。作为论文的读者，若是想验证这些数据，仍然可以到这家数据图书馆去获得。

王素的论文由于数据翔实、分析透彻，得到导师的好评，并最终发表在一个著名经济学刊物上。



第 4 章

进行数据革命

大数据成为推动经济转型发展的新动力。以数据流引领技术流、物质流、资金流、人才流，将深刻影响社会分工协作的组织模式，促进生产组织方式的集约和创新。大数据推动社会生产要素的网络化共享、集约化整合、协作化开发和高效化利用，改变了传统的生产方式和经济运行机制，可显著提升经济运行水平和效率。大数据持续激发商业模式创新，不断催生新业态，已成为互联网等新兴领域促进业务创新增值、提升企业核心价值的重要驱动力。大数据产业正在成为新的经济增长点，将对未来信息产业格局产生重要影响。

国务院《促进大数据发展行动纲要》

4.1 数据用于决策支持

4.1.1 数据分析需要统计而不是检索

在信息时代已积累了很多数据，成熟的数据管理主要是用关系数据库技术处理结构化数据。随着大数据技术的发展，面向非结构化数据的关系数据库技术及分布式数据库技术逐渐成熟。

数据的应用基本有两种形式：一种是数据的检索，它的特征就是不管数据量有多大，我们只找需要的数据；第二种形式是数据的统计分析，它的特征就是对单个数据并不感兴趣，主要是一个数据集表现出来的总体情况。

在数据分析中，数据的准确性并不太重要，因为一个大的数据集中即使个别数据出现异常，是不会影响到合计或者平均值的。

现在数据的大量应用都是一种特定形式的数据库检索。比如说搜索引擎，理论上可以搜索在网络上的所有信息，但我们常常关注的是符合搜索条件的其中一个或者多个数据。同样，许多软件提供的检索功能、检索的目标也是得到所需要的单个数据或者一系列个体的数据，这种应用在数据时代并不是主流技术。

数据时代的主流技术是统计分析技术。因为数据分析主要是找出它的统计规律而不是单个数据的内容。

比如，有一个人想出租一套房子，市场上可能有100个人有租房的意向，假如这个人能够看到这100个求助人的详细信息，那他应该如何利用这些数据得到尽可能高的租金呢？

按照数据库检索思维，他应该从数据库里找到出价最高的那个人，拿到可能得到的最高租金。但当他找到这个人的时候，有可能这个人已经租

了其他房子，原因是恰好遇到一个做中介的朋友，而这个朋友手上正好有套房，与出价无关。找第一个最高价的人没有成交，找次高价也有可能各种因素没有成交，通过几次这样的尝试以后，这个人觉得单个数据价值不大，因此放弃通过数据检索来寻找高租金的方法。

实际上，最佳的方法是他对这些数据进行分析，发现出价高的人是哪些职业，在哪些区域上班，将出租对象瞄准这些人，而不是去找单个的最高出价者。

找到规律后，就可以有针对性地发布出租信息。比如发现某个高档办公区或者金融行业从业人员租房出价比较高，那么只要在这个相关的区域或者相关的圈子里发布信息，就能找到出价比较高的人，虽然不能做到出价最高，但它的效率是最高的，从而达到资源最佳配置的目的。

4.1.2 数据通过辅助决策产生价值

数据自身不能产生价值，需要有一个转化的过程。转换过程是从数据变为信息，信息影响决策，决策产生价值。在一个特定环境中，通过对数据的解读产生信息，信息与环境有关，相同的数据在不同的环境中可能会被解读为不同的信息。

人们无时无刻，不管从事什么活动，无论是个人行为还是集体行为，决策无所不在。

与决策相关联的行为有大有小，大的决策如投资数亿美元建立或收购一个企业，小的如几点开车出门或走那条路。一个重大决策常常决定了很多资源的配置和价值。

决策的影响有大有小，一个大的错误决策投资可能损失几千万元，一个小的错误决策可能只浪费一个人几分钟。一个飞机驾驶员的错误决策，严重的会导致飞机失事。

决策实际上不是一个凭空的行为，而是一个人根据掌握的信息做出的决定，这种信息的准确来源就是数据。对一个飞机驾驶员来说，飞机内部设备的运行状况和外部的气候数据是他做出决策的重要依据。

我们每天的交通出行，对外面交通拥堵情况的了解是我们做出决策的主要依据。

能不能得到最全面准确的数据，这些数据能否被大脑接受，这是数据时代面临的主要问题。

数据分为实时数据和历史数据，一般认为实时数据更有价值，因此对实时数据关注更多，但实际上历史数据价值更大，只是开发利用的不够。

比如，现在交通要道上有一些交通拥堵情况的指示牌，用颜色指出前面路段的拥堵情况，绿色就是畅通，黄色就是有一点拥堵，红色就是很拥堵，司机根据这些数据决定直行还是绕道，这种数据特征：一是通过实时采集的数据；二是通过可视化的方式显示在路牌上，让司机非常快速地获知前面路段或相关路段情况。实际上，这些指示牌效果还是有限的，毕竟车子已经出门。驾驶员本来可以有更多的选择，或者早点上路或者晚点上路，甚至改天上路，等到了路上只能选择改道，选择很有限，而出门一般选的路肯定是最近的路，绕路的话虽然时间会节约，但燃油成本会增加。如果能预先做决策，就需要不仅仅收集实时的数据，更重要的是历史数据。可以通过历史数据判断每天下午 4 点半到 5 点半下班高峰时间路上肯定是堵的，这样的话可以选择早走或者晚走，所以历史数据比实时数据更有价值。

信息通过提炼规律可以转换为知识，有人认为是通过知识辅助决策的。显然决策离不开知识，决策是有知识的人利用信息做出的选择。决策者的知识可以来源于信息，但这不是决策的前提，因为更多情况是利用老知识，只有在做出多个决策后才能增加新知识。

4.1.3 两类完全不同的程序

很多程序员都没有意识到，世界上竟然存在两类不同的程序。大家都以为，程序都是用 Java、C# 等语言编程，用关系数据库管理数据。他们对有的项目很长时间不能收尾，客户需求总是难以满足感到困惑。

如果大家思考一下，这种客户需求经常变化的情况是否发生在项目最后的报表阶段，如果是，那就不是你的问题了，因为这个问题只有用另一

类程序才能解决。

现在开发出的程序很多，从事程序员工作的人很多，但做的都是同一类软件——事务处理软件，而报表属于另一类被称为决策支持系统的软件。

在涂子沛的《大数据》[7]一书中，描述了决策支持软件的发展历史。1947年，即人类第一台计算机问世的第二年，卡内基梅隆大学的赫伯特·西蒙开始了决策支持系统的研究。20世纪70年代，麻省理工学院的研究人员第一次提出，决策支持系统和运营信息系统截然不同，必须分开。1988年，IBM公司的两名研究员提出一个新名词：数据仓库（data warehouse）。1992年，比尔·恩门第一次给出数据仓库的清晰定义和操作性很强的实战法则，被誉为“数据仓库之父”。1996年，拉尔夫·金博尔提出“数据集市”（data mart）。1993年发明关系数据库的科德详细阐述了联机分析（OLAP）的定义。2000年以后，决策支持系统的理念和架构才完全成熟，很多主流的软件公司，如Oracle、IBM、微软、SAP通过自主开发和并购推出各自的称为BI的产品。

决策支持系统的技术包括ETL、OLAP、报表、可视化、数据挖掘等，无论开发思路、开发使用技术和工具都和一般事务处理程序不同，所以它是完全不同的另一类程序。

4.1.4 传统商业智能模式的沦落

虽然商业智能在2000年以后理念和架构才成熟，2010年进入高潮，出现大量的并购交易，大型软件公司都通过自主开发和并购建立自己完整的产品线，但到现在，创新已经停滞，市场大幅下滑。经历从商业智能到商务分析的改名，却逐渐走入黄昏，很多从业人员已经改行。

Gartner 2016发布的BI和分析魔力象限[8]印证了这种趋势。在这幅图中（见图4-1）可以看到，2016年，传统BI厂商集体沦陷，全部被驱除出了领导象限。IBM、SAP、SAS、Microstrategy等无一幸免，Oracle甚至已经完全消失。

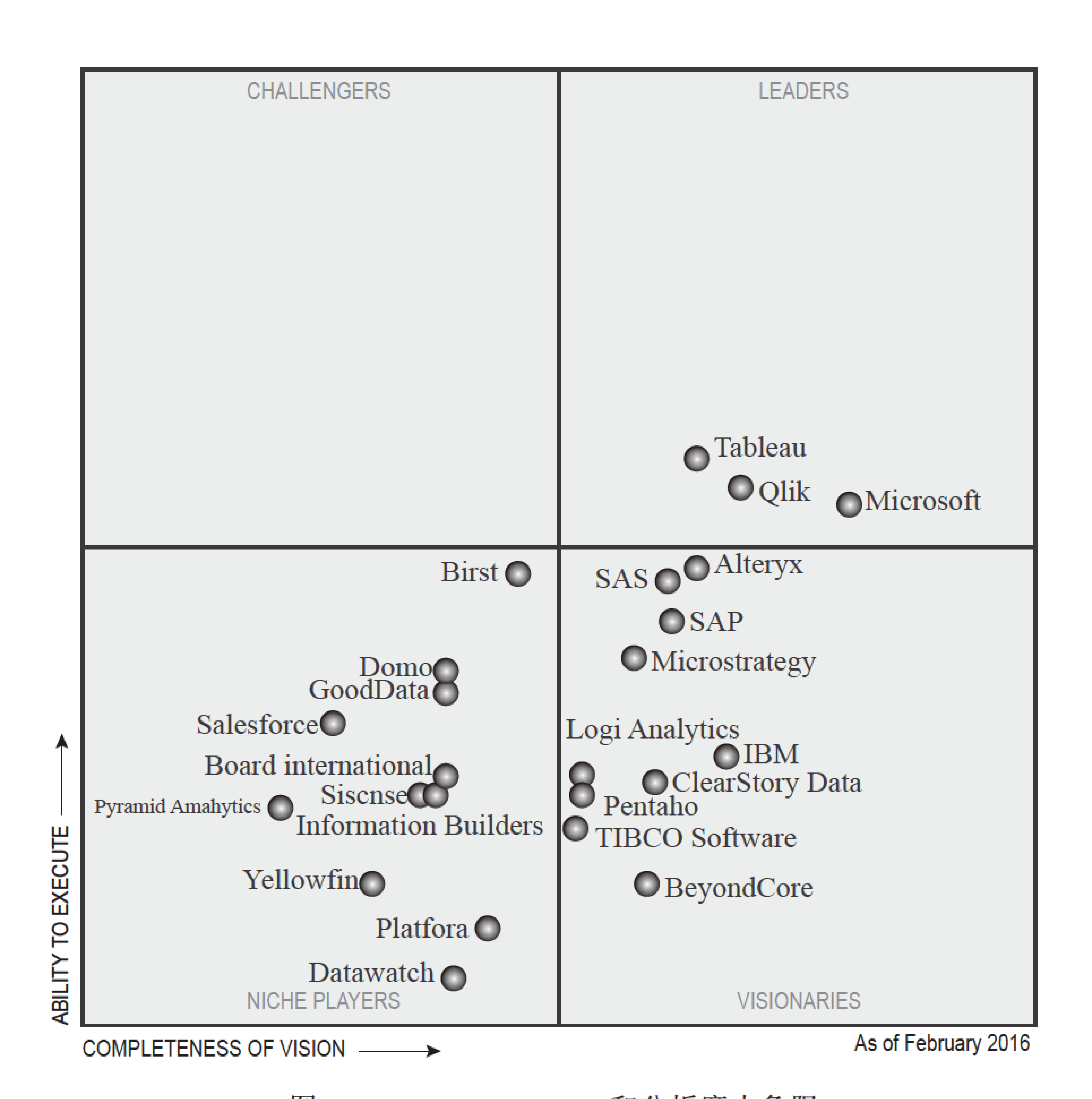


图 4-1　Gartner 2016 BI 和分析魔力象限

传统 BI 厂商的沦落有好多原因，比如来自并购的不同产品集成度差，相关联问题分割成不同层面，用不同方法、不同软件产品去解决，但核心问题是无法满足客户需求，用开发事务处理软件的流程去开发决策支持系统。比尔 • 恩门的书 [21] 中明确提出，开发决策支持系统要先有程序，后有需求，但实际开发中还是要求客户先提需求，因为客户无法提出需求，或需求不全面将导致项目完成后新需求无法满足，会影响客户的满意度。

如果没有解决好需求问题，商业智能发展仍然死路一条。Gartner 推崇的以业务用户为中心的自助分析工具或平台，仍然没有解决这个问题，只不过把需求这个难题由 IT 人员交给了业务人员。虽然业务人员相比 IT 人员更了解需求，但仍然有许多当前没有想到、没有遇到的需求。此外，由业务人员使用的自助分析程序，加大业务人员素质要求、增加工作量，会

提升企业的用工成本，而且很多工作成果难以共享。

传统BI厂商的技术和产品还是有价值的，如ETL和OLAP服务器技术，但开发思路需要改变，要能基于无需求进行开发。

4.1.5 像鹰一样看数据

在我们拥有数据以后，如何让数据发挥作用？

数据发挥作用的方式有两种：辅助决策和数据驱动。数据驱动指在业务运作流程中以数据结果为运作目标，以关键数据为触发方式，借助计算机相关技术结合企业内部流程和机制形成数据一体化的工作流程。

数据驱动不需要人为干涉，但这种方式带有明显的局限性：第一，它可能是一种个性化的开发，成本比较高；第二，它开发的目标比较单纯，不具备通用性。

相对数据驱动，辅助决策能够产生更大价值。辅助决策要求能看懂数据。那么，理想状况下，我们应该怎么样去看数据呢？

人们都想像鹰一样能在天空翱翔。如果从鹰的视角来看，觉得自己非常自如，能飞在很高的天上，看到很广阔的地域。假设把视野所及地域铺满数据，或者想象这片地域是由大数据组成的，那么飞得越高，看到得数据就越多。

鹰可以看得比较宏观，但它又可以随意地调整高度，缩小在自己视野里的区域，甚至在发现一个猎物时，它可以从很高的天空俯冲下来直扑地上的一个点。鹰眼非常符合我们看数据的要求。

在看数据时，宏观和微观怎样融合，怎样自如地切换？从鹰的行为可以出来，宏观和微观各有所长，没有宏观的视野它很难在一个广大的区域里找到猎物，如果没有微观捕获猎物的技术它就无法获取猎物，所以两种技巧都需要。

实际上，在网上应用地图的时，我们已经拥有类似的视野。我们可以在网页上打开一份世界地图，然后用鼠标任意缩放，通过鼠标滚轮将它缩小到一个国家、一个城市甚至一条街道。这种数据组织方式的特点是，虽

然拥有非常大的数据集，但不会一下子被这么多数据包围，总是可以看到一个小的，人能够接受的数据集。比如在世界地图上，只显示一些国家的名称和主要大城市的名称，明显只是一个小数据，但放大到一个国家以后，可能就把这个国家的一些详细数据显示出来，展示这个国家重要地区的省份或者城市，再缩小地图，才会显示一些更小的街道。

同样，在利用普通数据时，虽然不能像地图一样按区域或城市的级别确定数据是否显示，但可以通过数据的合计和平均值来达到类似的目的。也就是说，可以从比较大的数据集的合计看到组成大数据集的相对比较小的数据集的合计，最后看到明细数据。这就是为什么我们把交互式可视化无方向数据挖掘这项技术命名为“鹰眼”技术的原因，通过这种技术，利用数据的钻取操作，可以使你具有鹰一样的眼睛，在大数据中找到你所需要的信息。

4.1.6 数据一致性不是分析的先决条件

在数据分析时，会发现应该相同的数据出现了不一致，比如合同数和发货数不同，生产成品数和入库数不同。出现这种情况一般有两种原因：业务流程控制不严和不同信息系统的主数据定义不一致。

出现数据不一致，是否说明信息化水平不够，还不具备做数据分析，或者决策支持系统的条件呢？

实际上，决策支持系统和事务处理不同，即使数据不一致，也不影响使用。而且，将两种完全不一样的数据放在一起比较，就能通过数据分析发现管理中的问题，并找到问题出现的原因。

不像事务处理系统，数据前后必须对应。在数据分析中，无论是同一个软件中的不同数据，还是来自不同软件的数据，都可以放在一起显示，并不强迫数据必须保持一致。比如，企业销售中有多个环节：合同、订单、发货、出库，理论上来说这四个数据应该一致，而实际业务操作时并非如此，经常出现没有订单就发货的情况。如何保持数据一致，是公司管理规定的执行或事务处理软件流程控制问题，与数据分析无关。而且把四个数据放

在一起比较，管理者会很明显地看出数据的不一致，会根据企业管理实际追责或调整。到底数据的不一致合理与否，是人为录入错误，还是适应市场的无奈之举，是否需要改进，都是一个管理决策问题，应交由业务部门而不是数据分析人员处理。数据分析人员只是提供一个决策的辅助系统，也就是把数据是否一致的信息准确地传递给决策者，而让决策者判断其正确与否。

从另一个角度看，任何一个信息系统或企业管理制度都需要不断地反馈，比如按公司规定出库必须有订单，但老客户没有订单发不发货？因此，规定和实际总有矛盾，上级要求与下级执行也会有出入。

假设在数据分析里，把订单数据和出库数据及时反馈给决策者，就可以明显地看出管理制度实际执行的差异。因此，通过数据分析建立反馈机制，让业务部门决策者看到自己的决策是否被执行，才有利于具体工作的落实。

如果数据分析需要等待数据的一致性，而数据的一致性问题由于没有得到及时的反馈、检查和督促，很长时间不能解决，那么数据的一致性将永远达不到应有的水平。

所以，数据一致性应该是业务部门在事务处理中不断调整的结果，而不是数据分析的前提。

4.1.7 从数据比较中发现价值

对于事务处理软件开发而言，数据的一致性很重要。比如说一个企业ERP 软件中合同的订单数量和仓库的发货数量，以及最后收款金额应该有对应关系，如果流程上出现问题少收款，就说明企业在管理上存在问题。

在数据分析中，对不同系统的数据如何一致，是一个困扰 IT 人员的难题。MES 的产量数据怎么和 ERP 的入库数据相一致？因为处于完全不同的信息系统、不同的数据库中，要一致比较困难。如果等数据都一致，再做数据分析和决策支持，显然对基础数据的要求就太高了。

要求数据的完全一致，是一个事务处理的思维而不是一个决策支持的

思维。这个思维的主要缺陷就是认为信息系统是完美的、可以在逻辑上完全一致的，完全排除了人的作用。

实际上，数据分析技术的核心不在于数据的一致性，而在于要让这些数据被人的大脑所认知。比如，销售合同数和仓库发货数应该一致，如何防止不一致的情况出现呢？按事务处理软件的流程控制思路，这种一致性应该由软件控制。但是，从数据分析技术来说，应该由人来控制，即把合同数和出库数同时展示给决策者，由他来判断这里面的不一致是否合理。当然，他看到的数据不仅仅是总额的区别，可以按客户、存货的品种去分析哪里不同，甚至按时间对两种数据的差异进行比较。

同样，对来自 MES 的产量和 ERP 的入库数的差异，也是把相关的数据展示出来，让决策者去分析。这样的话，对数据源的数据一致性要求大大降低，充分发挥了人脑的作用。这样处理有利于决策者通过发现数据的异常来去不断地调整，也许最终的目标数据是一致的。但这一致的过程不是说技术人员层面就能搞定的，而是需要在决策人层面提出要求，因为这个不一致可能涉及业务流程的重新设计，涉及不同业务人员的调配、工作责任心和录入数据的及时性。这些问题不是 IT 人员能够处理的，需要从公司的角度协调多方面的力量共同解决。只有达到这样的目的，决策支持系统的目标才能实现。

4.1.8 保障决策者的决策思维流

笔者有一个大学同学是杭州一家汽车配件企业的总经理，笔者曾在他们公司和他共同研究了利用 ERP 数据的决策过程。

他们用 ERP 已经好多年，还是当初利用政府补贴购置的。笔者坐在他的办公桌前，看着他操作软件，调用菜单查询。他查到公司当年销售额数据，发现销售额只有当年、上年同期、同比增加额的数值，没有计算同比的百分比值，更没有图形。他想进一步了解销售额的分布，需要打开另外一个菜单才能看到按客户或产品的销售额分布，找到销售额增加或减少的原因。

于是，我们发现这种查询，导致一种可称为决策思维流的中断问题，

决策支持系统的一个重要的功能就是要满足管理者决策思维流的需要。

大家知道，在电影中有一种技术叫意识流，这个是电影的一个表现手法。在电影上可能看到一些跳跃的画面，表面看没有很明确的时间关系，但它符合人类大脑在思考问题时意识不断跳跃和流动的过程，所以在电影手法中称为意识流。意识流在刚出现的时候有些观众难以理解，但随着几十年下来不断地普及，观众对电影意识流的手法慢慢就熟悉了。

同样的，在我们决策者思考问题的时候也有一个流，可称为决策思维流。一个决策者在发现一个问题后，肯定需要对这个问题做进一步的思考。比如说，企业在发现销售额下降以后，就会想销售额下降是什么原因引起的，是哪个区域销量下降了，还是哪个产品的销量下降了。如果一个区域的销量下降了，会想是这个区域里所有客户的采购量都下降了，还是某个客户的采购量下降；如果是某个客户下降，要知道是他采购的所有产品量都下降，还是某个产品的采购量下降。如此这般，就会形成一个思维流。

一个决策支持系统应该具备什么功能才能满足决策思维流的需要呢？

首先，发现问题都是随机的。可能是自己分析数据的时候，看经营数据的仪表盘时发现的问题，也有可能是下属汇报时的问题。可能是销售的问题，或者是收款问题，也可能是库存问题。由于问题都是随机的，所以当问题出现时，要能及时提供相应的数据。

其次，要为思维流的流动提供进一步的操作功能，比如可以进行数据钻取、变化维度查询。

最后，数据显示的速度要非常快，要跟上决策者思考的节奏。比如说当这个区域销量下降之后，要了解是哪个客户下降了，就要马上看到这个区域所有客户销售的同比数据。

如果由于技术的限制，当他想到下一个问题的时候需要别人协助，思维流就会中断。比如，如果一个区域客户的销售同比没有计算出来，并且没有图形化，如果他需要业务部门或者秘书来做这个图形的话，这个问题他今天就无法思考下去，他也不可能再去提下一个问题。

如果数据统计或显示很慢，比如要第二天或者要过一个小时才能提供数据，这个问题也无法按照思维流进行下去，问题可能就搁置下来。等下

次有了这幅图，他回想起这个问题，发现有一个客户确实下降了，再去找哪一个产品下降的时候又拿不出数据。如此这般，他就无法按照这种方式去思考。

我们可以想象，如果按照这个思路一直想下去就是这个结果。如果你打断了思路，可能就得按另外一个思路去思考，很难回到原来的思路上去。

所以作为决策支持系统，它的目标就是要满足人们对决策思维流的需要。应该说，如果软件不行要做软件，硬件不行要升级硬件，这样才能真正满足决策者的需求，充分提高决策的效率。

4.1.9 建立基于可视化数据的指挥室

一个组织无论是政府部门还是企业都应该有一个指挥室，就像军队的指挥室一样。现在有些政府机构已有类似指挥室的地方，有了指挥室的雏形，但基本以视频监控为主。视频监控提示的是实时视频信息，缺少数据信息。在一些企业，商业智能公司开发了数据可视化系统，在上面显示公司的实时经营数据，类似于军队指挥室里面布置的是沙盘、地图，见图 4-2。

一个指挥室应该以决策支持系统为主，而决策支持系统应该以数据为主，数据除可监控最新经营状况外，还可观察历史数据的变化，多维度对数据进行分析。

指挥室不是一个监控室，监控室作为组织的日常事务处理部门主要应对突发的事件。指挥室应该不是用于程序化的突发情况的处理，而是要做一些对公司有比较深远影响的战略决策，这些决策更多要依赖历史数据去分析问题，并且要对决策的长远效果做出反馈和评估。

指挥室有两种布置：一种为影院式，即大屏幕放在前面，后面椅子直对屏幕，可称为老板指挥室，主要适合以老板为核心的管理团队来进行指挥决策。还有一种是围桌式，类似会议室，适合开董事会和办公例会，对应有大屏幕，屏幕控制可以采用触摸屏。现在微软推出来的 Surface Hub 是一种大屏幕触摸屏，上面可以实现视频会议、白板及 PPT 展示等功能，显示决策支持系统的话可以直接在上面操作。如果人少的话，可以在

Surface Hub 上面直接看，它有 55 英寸和 64 英寸两种规格。如果还嫌不够大，可以把画面投影到大屏幕上，让两者同步。由于决策支持系统支持决策思维流，很多的问题在这个指挥室里都可以解决，大大提高了决策效率。



图 4-2　招商局集团的可视化平台

4.1.10　组织的决策支持流程

这里的组织包括政府、教育机构、医疗机构、企业等。

组织的规范决策一般可以分为几个过程：①数据的收集过程；②数据的展现；③数据的分析；④决策；⑤决策结果反馈。

一个组织要进行决策，需要掌握大量的数据。就像战争中，要通过侦查员、飞机、卫星去搜集大量的情报，显然情报搜集得越多、越精确、越及时，对决策的作用也就越大。所以数据搜集是正确决策的必要前提。

同时，决策对数据有几个要求：①拥有数据；②看到数据；③看懂数据。

每一个组织都有很多的数据来源：①人工直接录入的数据；②自动记录的人员活动数据；③设备运行产生的数据。

在企业里面，员工操作 ERP 软件录入的订单的数据、出库的数据、收款的数据等，是人为录入的数据。网店的用户点击操作的日志、生成的订单、消费的金额，员工的考勤打卡等，是自动记录的数据。设备运行产生用电、设备故障、报警、检修等，是机器运行产生的数据。

另外，在医院里，门诊挂号、医生开处方、病房输液操作、手术等会产生数据，一些医疗设备也会自动产生数据，比如说 X 光机，开动时间、运行时间等。政府也有行政审批、报警的数据。金融机构，有存款、取款等数据。

要方便数据搜集工作，首先，设备要智能化，设备上都要有传感器，它的运行状态都有数据输出；其次，这些数据能通过网络送到一个服务器中存起来。

现在有些企业为了节约成本，虽然购买的是数控设备，但设备上数控模块是选购的，就不选了，所以这个设备还不算智能化。有些企业虽然选购了数控模块，但没有联网，所以数据也不能存储，更不可能分析了。

数据的搜集只靠设备智能化是无法全面完成的，有些事还需要做人工处理。人工处理需要用到一些事务处理的软件，比如制造执行系统（MES）。如果现成的软件包适用的话可以采购，不适用的话需要去定制开发。

这些数据都全了以后，可能存放在不同的系统中，这个时候需要开发一个数据仓库，把这些分布在不同服务器上的数据集中到一起，然后把数据仓库的数据通过建模变成小数据，最后把这些小数据可视化处理后发送给领导。

领导在自己的办公室里，可以看到所有数据源的数据并且对这些指标进行监控、分析。可以制作一个仪表盘进行监控，发现问题以后通过多个维度的分析、逐级的数据钻取来查询数据，找到发生问题的人或设备。

可以设定一个理想目标：如果所有设备里面有一台停机了，这个停机的信息应该能反映在老板桌面最起码的一个指标上，比如入库数量。因为一台机器停产，出产的数量就少了，入库的数量也会减少，继而在决策支持系统上就能看到入库数量这样一个指标的变化。一个企业如果有一百台、一千台机器，一台机器的停产显然是不会影响到入库数量的总额，但可以

对产品入库数据再给出一个环比指标来解决这个问题。影响环比变化的设备生产数量虽然很少，但是一停产，和昨天的环比就会产生很大变化，这个环比的变化会给老板一个异常发生的提醒，他通过数据钻取，可以逐级追踪到具体的分厂、车间，最后定位到这台设备。决策者会看到这台设备因故障停产了，同时可以评估这台设备的停产对其他经营指标的影响，做出相应的决策。

4.1.11 宏观和微观的融合

宏观和微观指对同一个事物不同的观察角度。宏观一般比较粗但关注的范围比较广，而微观比较细但关注的范围比较窄。

在管理中微观和宏观常常是互相排斥的。一般来说，领导的层级越高他看到的越宏观。如果最高领导者太关注微观就会缺乏宏观的把控，在对整个组织的发展把控中就会出现问题。

在美国击毙拉登这个事件上，通常总统只需要关注宏观的决策，就是要不要击毙拉登，至于如何去击毙拉登，在什么时间、什么地点击毙拉登，应该由下面具体的执行者决定，不但不需要总统来处理，甚至不需要五角大楼的高级将领来处理，只需要由现场的海豹突击队的队长来处理就行。

但从新闻中看到，不仅奥巴马全程参与，而且通过海豹突击队的队员的头盔看到现场的视频。

这样就带来一个问题，我们传统观念中管宏观的人不需要关注微观，到底是基于人的精力限制还是技术的限制。精力限制指一个人过多关注了微观就没有精力关注宏观，若是太关注微观就会影响宏观判断。技术限制是由于技术的局限，导致人的活动范围和搜集信息详细程度的能力受到现在而无法看到微观的信息。比如中国古代的皇帝是通过奏折或者是通报来获取下面的信息，这种奏折由于是手工书写的文字，所以不可能非常的详细，而且这些事情发生在全国各地，皇帝也不可能亲临现场，也就是说他没有条件获取微观信息。

随着技术的发展，电视直播成为一种可能，也就是说作为中央领导可

以通过直播看到现场的情况，完全可以掌握到相对微观的信息。交通工具的发达到达现场也成为一种可能。

因此，可以得出这样一个结论，虽然从管理的权限和层级来看，为了防止管理的混乱或者政出多门，指令应该是逐级下达、逐级执行，作为上级来说过于微观的命令会干扰基层领导的工作。但从信息的采集来说，微观和宏观却不应该有很大的差异，毕竟有很多宏观的信息是来自微观的综合。微观事件的发生会有一个时间性，也就是在指定的时间或者时间段里微观的突发事件总是不多的，作为一个领导来说完全有时间关注微观信息。

从来没有一个固定模式规定高层领导只需要关注宏观信息，而不需要关注微观信息。为了正确决策，一个高层领导既要掌握宏观又要掌握微观，关键是如何在最短的时间花最少的精力获得最详细的信息，怎样从宏观上尽可能快地发现微观的异动，把关注点转移到对宏观有影响的微观事件上，并且迅速掌握详细的情况，以便于协调事情的处理，并得到处理结果的反馈。

作为一个技术人员或者一个基层领导，没有任何权力阻止上层领导获取微观信息，反而要保证他获取到最细粒度的数据，这样，这些信息可以根据宏观管理的需要去按需索取。既然宏观的管理影响面比较广，它的任何一个决策都会产生非常大的价值或者损失，所以即使这些微观数据一年都用不到，但只要用过一次，它的价值就完全可以抵销它的成本。况且在现在的技术条件下，这种工作并不是专门为宏观决策者准备的。也就是说，宏观决策者和微观执行者是共享数据的，只不过需要在技术上将数据打通，所以这里面也并没有特殊的成本存在。

在经典的商业智能开发模式中，把宏观和微观功能人为地割裂开来，把系统分成决策层、管理层和执行层，不但开发的内容不同，而且使用的工具也不同，比如决策层用仪表板，管理层用 OLAP，执行层用报表，不能实现宏观和微观的融合，影响了作为决策支持系统的价值。

4.1.12　用过度设计满足任意需求

现有一种概念，称为过度设计和过度服务。其本质是在一个产品的设

计上或者是在服务上超过了客户的需求，即过度了。

过度设计是在产品同质化越来越严重以后，为了突出差异性增加竞争力才出现的。有些公司在设备上不断增加新功能，这些功能有些确实提升了产品的效用，但有些没有作用，设计这些没用的功能就称为过度设计。

通过增加产品的功能，而增加产品和竞争对手的差异化，是过度设计的初衷。但是，也有些公司过于沉湎于细节，究其原因是在大的方面上没有创新，只能集中在细节上进行创新，从而导致过度设计。在日本，很多公司把过度设计和过度服务当成一种主要的竞争策略。

就数据分析而言，过度设计是面向需求无法描述的现状的一种应对策略。客户能提出的需求，或者能感觉到需求，只是总需求的一小部分，现在不需要不等于以后不需要。如果一个系统只能满足现在的需求，那么只要一有哪怕一点小的新需求就重新开发升级，显然是不能被客户接受的。

现在很多公司提供的自助式开发工具，实际上是一种推卸责任的做法。自助设计是个非常复杂的工作，就像用 Photoshop 软件可以做出非常好的美工作品，但不等于有了 Photoshop 软件之后随便什么人就能设计出同样效果的作品，这和操作者的美术素养和设计眼光都有很大的关系。同样，客户要做好自助开发不是件很容易的事情。

所以，相对于客户当前的需求来说，要满足他未来的需求时，必须要进行过度设计。所以"过度"，相对于现在的需求来说是过度的，但面向他的总需求来说实际上还是不足的。

使用过度设计方法开发，要求开发者的经验要高于所面向的客户。在事务处理软件系统开发中，很多跨国公司从世界范围内收集一些好的实践案例，然后和客户的需求结合，导致这些系统开发出来比仅依赖某个具体客户的需求更加完善。但在商业智能系统开发中，并不能复制类似的成功。

因此，在决策支持系统的开发中，靠简单拼凑经验来完善需求已被证明此路不通，原因就是需求的范围比想象的要大，而且大得多。

归纳法行不通，只有用演绎法。可以根据模型推导需求，这样得到的需求相对而言科学性和理论性都比较强，才能满足客户的需求，如此多的功能对客户来说就是一种过度设计。

面对过度设计的系统，客户一开始会有些不适应。定制系统都是根据客户的需求来完成的，客户需求也是考虑了好长时间。经过了长期的酝酿和思考，开发者主要工作是理解这些需求，所以思路是跟在客户的后面的，程序开发处理后客户理解非常容易。但经过过度设计，客户完全没有准备，一开始时会抱着排斥的态度，在慢慢使用、熟悉以后才会逐步接受。

4.2　建立数据模型

4.2.1　存储数据的数据仓库

一个组织内部通常有许多信息系统，不同信息系统采用不同的数据库，而且由不同的公司在不同的时间开发和实施，甚至可能有不同的版本，更甚者中间可能更换过不同的供应商。

如果要从整个组织的角度来利用数据，一般需要数据仓库。这种数据仓库又称为企业级数据仓库，简称 EDW，区分于常常和数据集市混淆的那种数据仓库。

建立数据仓库的第一个目的，是把不同系统的数据库的不一致性去掉，也就是用同一个数据库来替代不同的数据库。将原来可能用 SQL Server、Oracle、MySQL 等不同数据库保存的数据，保存到一个统一的关系数据库中。一般都用普通的关系数据库做数据仓库数据库，比如 Oracle。也有专门的数据仓库数据库，比如 Teradata。如果用 Oracle 作为数据仓库数据库，就必须把所有的在其他数据库的数据统一到 Oracle 数据库中。

第二个目的，因为不同的数据库中表和字段有不同的定义，有些定义数据库的数据字典已不存在，而且时间长了，开发人员可能也找不到了，从而导致数据库中数据再没有人能看懂。在这种情况下，如果把数据移植过来，因为数据仓库里数据定义比较一致，大家就能看懂了。

数据仓库和数据集市的目的不一样，不是为了最终的查询，而是为了

要保存原始的数据。

目前数据仓库做得比较好的是银行。根据银行监管部门的要求，交易数据必须保持三年到五年，所以银行把几十个业务系统的数据统一保存到数据仓库中。银行一般采用 Teradata 的服务器和数据库来建立数据仓库。

数据仓库数据库虽然从架构上来说和关系数据库很像，但在数据库设计上需要专门的模型，保存的数据包括数据是来自哪个系统、什么时候抽取过来的等这些数据来源信息。即使有些数据信息存在不一致，比如一个人有不同的住址，也要保存原始数据，以便以后分析时能查询到这些区别，需要时再对这些数据进行甄别。

实际上，随着信息系统的发展，公司的系统不仅是多系统共存的问题，还有替代的问题，所以历史数据的保存也非常重要，数据仓库是保存历史数据很好的工具。

然而，现在数据仓库的普及还不够，原因除投资比较大以外，还有就是它的价值还没有体现出来。比如银行的数据仓库，更多的基于数据备份的目的，并不是支持用户进行数据分析。Teradata 也推出了探索式数据分析工具，叫 Aster，但它的查询语言是 SQL 语言。就是说，必须懂 SQL 语言才能用 Aster，所以使用对象只能是专业人员，而不是一般的业务人员。正是由于直接应用的缺乏，导致大家对数据仓库的建设不太重视。

另外，成本也是一方面。现在数据仓库硬件软件都是一体化的，无疑成本很高。即使是银行，也有成本压力。现在的 Hadoop 等大数据技术的迅猛发展，就是为了解决这种成本高企的一种替代方案。相信随着软件硬件成本的降低，数据仓库会得到越来越多的应用。

图 4-3 是企业数据仓库的四层模型，一般书上只有三层模型，把企业数据仓库和数据集市并列，两者只要一个即可。

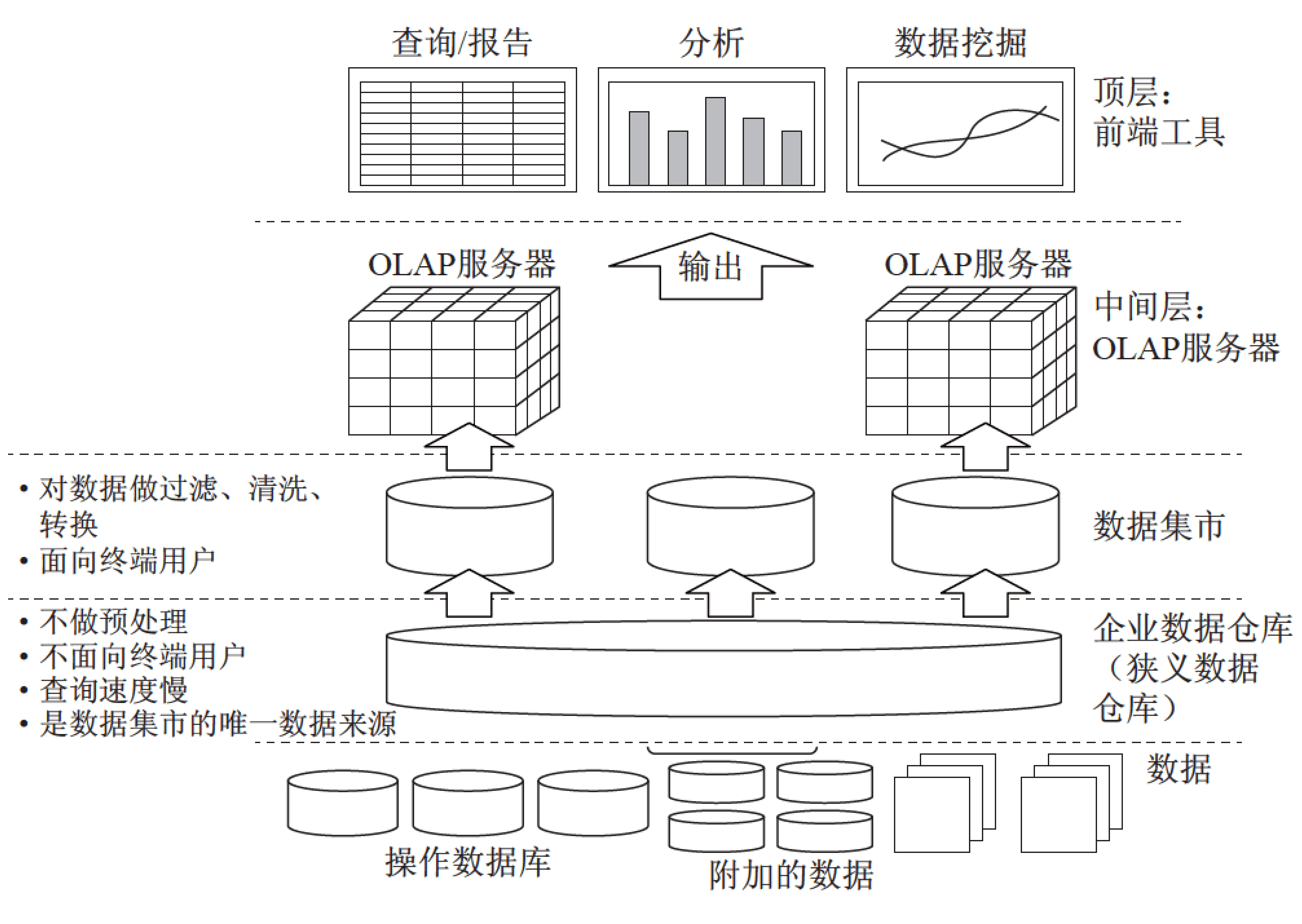


图 4-3　企业数据仓库的四层模型

4.2.2　可以推导需求的维度模型

现在中国的软件开发从哲学上来说是经验主义，而不是理性主义。经验主义完全根据客户的需求出发，客户需要什么功能就开发什么功能。

在软件编程技术越来越普及以后，软件公司不去技术上寻求创新，而只是关注行业经验，认为只有掌握行业经验，软件才有价值，也就是说，他们认为只有凝聚了经验的软件才有价值，纯技术是没有价值的。

鹰眼技术的核心是维度模型。从使用角度来看，采取从小数据看大数据，然后逐级钻取的方式，从一个数据的合计开始，按照一个预先定义的层次结构逐级钻取到最细的一级，直到钻透到明细数据。钻取时采用可视化方式，每个小数据都不是以数字的方式，而是以可视化的方式来显示。钻取的工作也在可视的图形化上交互式进行，符合人们的使用习惯。

虽然维度模型出现以后，很多人也在使用，但可能更多局限在 OLAP 建模的时候。只在建模的时候用到它，在展示数据时就丢弃了。鹰眼技术

的主要特征是维度模型的使用贯穿数据仓库数据库的建立、OLAP 建模及最后交互式可视化的界面设计全过程。

使用维度模型，关键的是把它作为人们认识和使用数据的一种方法，而且是一种通用的方法，并不受人们经验的影响。也许从经验的角度来说可以对这个模型进行微调，做一些取舍，但总的架构应该完全基于这个模型。

当用户面对这个模型，开始时会非常不适应。因为它明显的是一个过度设计的产物，许多功能超出了人们的想象。一般来说，软件的用户希望根据自己的认识水平来逐渐增加软件功能，一下子接受很多的功能的话会非常不适应。但这种设计却能够满足用户的许多需求，包括现在的需求和未来的需求，他们的需求一般很难超越模型覆盖的范围。

用户的需求在不断的变化之中，可以分为现在的需求和潜在的需求。潜在的需求现在还没遇到，所以用户也说不出来。还有的需求是已经有这个意识，但无法描述出来。在做用户需求调研的时候，用户仅仅能说出现在意识到的，并且可以描述的需求。这就是许多定制软件在客户使用之后总是需要修改的原因，特别是在做报表时，问题更为明显。报表实际上是决策支持系统的一部分，因为使用报表的时候常常会遇到一些不确定的需求，在做需求调研的时候，用户只说出了部分需求，而在实际使用的时候，他们又会觉得报表不能满足要求，而软件开发者觉得客户的需求变化太大，无法把握。

现在很多的软件公司的解决方案是提供一个报表的自助设计工具。你需要什么报表，自己去定义。但是，一个报表工具的使用也不是简单的事，需要专人去学习使用，而且报表使用者和开发者也不是一个人，只是把供需矛盾从软件开发公司转移到企业内部的开发人员，并没有真正解决问题。报表真正的使用者可能是一些领导层或者是业务部门的负责人，他们在有需求的时候会交代技术人员去开发，而对技术人员来说，他的需求也是不确定的，可能今天要这个数据，明天又要那个数据。出现这个情况的主要原因是客户描述的需求只是总需求的一部分，类似它的一个子集，而客户实际的需求虽然有一部分会在这个子集内，但常常会超出这个子集。这个需求不确定问题在报表系统开发中是一个常态，如何应对呢？鹰眼技术是

采用的一种基于理性主义的、基于模型的推导，它不受客户目前需求的限制，可以推导出几乎所有的需求，为谨慎起见，我们的说法是——80% 的需求。或者从另外的角度来说，对客户是过度设计和过度服务，从表面上来看远远超出现有的需要。只有这样才能满足客户。

4.2.3 维度模型原理

在传统数据库中数据检索技术已经非常成熟。

数据检索就是给定一个或一组条件，找出满足条件的数据子集。检索条件对应数据库中数据表及表中某个或者某几个字段的值。比如显示一份上海市年龄大于 80 岁人员的名单，就是一个简单的检索功能。

如果有这么一个需求，要求显示上海市 80 岁以上年龄老人的平均收入或者是平均医疗费用，这种数据就属于数据统计。如果用普通关系数据库去统计的话，速度会比较慢。

为提高计算平均数或者合计数的速度，常常采用一个叫轻度汇总方法，也就是把可能需要的统计数据预先计算好，比如说在晚上进行计算，然后把它保存在另一个数据表中，这样如果需要查询的话，就不需要从原始数据表中读取数据再计算，而是直接从轻度汇总表中查出计算结果，从而大大提高数据统计的速度。

但是，轻度汇总的问题在于难以满足使用者的各种需求。因为你可能按几种条件组合去预先汇总，而他可能正好查询的是你汇总之外的一组条件。比如说，按照 80 岁以上年度收入合计做了预计算，但 70 岁到 80 岁的月收入平均值没有预计算，这样前面一个条件速度很快，查后面一个条件速度就会很慢。如果你想把它所有的条件组合都汇总，但这些条件组合会很多，比如说，汇总是按年度汇总，还是季度汇总、月度汇总、日汇总。汇总条件组合太少，查询命中率就很低。但如果都汇总的话，由于组合会非常多，会出现所谓的“维灾难”，根本无法做到。

另外，汇总计算需要的时间非常多，晚上可能由于数据太多，没有足够时间去计算，可能到了天亮还没有汇总完。如果一个月汇总一次，那么

数据更新过后，汇总数会来不及更新。还有一个问题，就是每个汇总都会占一定的硬盘空间，这么多汇总硬盘是否能放得下。

很多计算机专家实际上早就已经研究过这个问题，是商业智能或数据挖掘领域的一个课题。对于这类需求，需要建立数据模型，这个模型叫维度模型或星型模型，由OLAP Server去处理。虽然这方面技术已经比较成熟，主要的数据库软件公司比如微软、Oracle都有相应的软件产品，但实际应用还不太多。

这个技术的研究高潮应该在2000年左右，但现在掌握并运用这个模型的人很少，有两个原因：第一个原因是这个模型相对来说比较难，属于一个专业领域，除非专业从事商业智能开发的技术人员，一般软件开发人员不掌握。第二个原因就是商业智能技术到现在成功应用的还很少，从表面上来看是由于技术要求高、成本高，很多企业用不起，实质上是它的失败率很高。就是说，即使花了很多钱做了BI的企业，大多并没有达到它预期的目标。所以这个技术口碑不太好，推广成本很高。

BI项目实施满意度低的主要原因在于对需求的把握。它还是按照传统的信息系统的开发模式，需要客户单位提供详细需求。客户单位提不出需求，或者即使勉强提出需求，需求又会经常发生变化，最终导致按预定需求开发的软件不能满足后面的需求变动，从而导致客户的满意度下降，最后导致了推广不力。有关如何满足客户需求不确定的问题，在本书其他章节里有详细的描述。

那么维度模型的核心是什么呢？

现在的数据很多，不同单位信息系统的数据格式都不尽相同，用户对于数据查询的要求也不同。

表面上看，BI面对的需求是比较混乱的，难以找出规律。这个非常类似牛顿在发现万有引力定律之前，人们对看到的很多自然现象难以理解：为什么扔下一个东西不会飞到天上而是掉到地上。在牛顿的万有引力定律发现以后，人们会发现实际上所有东西的运动都受万有引力定律的限制，这样就很好地解释了世间各种各样的现象。

同样，维度模型也是找到了数据检索的规律，从而把看似没有规律的

需求归结为一个统一的模型。客户需求的规律就隐藏在维度模型中，只要按照维度模型对数据进行处理，用户的需求一般都能满足。

这个维度模型把数据分为两类：一种数据叫维度数据；另一种数据叫度量数据。维度数据常常是文字、日期类型，比如说客户名称或者销售日期，事实表数据都是一些数值类型，比如销售数量或销售金额。用维度数据做查询条件，一般来说不会跳出维度表范围，对事实表里面的数据可以进行汇总，比如说算合计、平均值、最大值和最小值等，这种数值表中的数值数据又称为度量值。

数据的查询是非常复杂的，为了对应这种复杂性，维度表需要进一步的划分，划分成层次结构。比如，日期是一个维度，包括年、月、日等，把这些数据库生成一个层次结构，这个层次结构里面含年、月、日三层，有了层次结构后，可以把日的数据自动汇总为月的数据，月的数据汇总为年的数据。每个层次结构里面包括多个级别，一个完整日期层次结构的级别可以包括年、季、月、旬、日。

在国外还有财政年度，有日历年度的年和财政年度的年的区分，同样月也有也有财政年度的月和日历月度的月的区分。在国内设计日期维度一般有两个层次结构就可以了，第一个层次结构叫年、季、月、日；第二个是年、周、日，大家知道周和月是不可以同时存在的，所以必须分成两个层次结构。客户查询数据不是按年查就是按月查，不是按月查就是按季查，或者按周查，反正查询不会突破这两个层次结构。只要按这两个结构进行数据处理，实际上也是一种轻度汇总，客户的需求难以逃出这个范围。

当然还有一些需求，对任意日到任意日之间的数据进行统计。这种设计对客户有一定的吸引力但实际用处不大，是一种偷懒的做法，以为这样就不要去研究客户的需求了。虽然客户会有这方面的需求，但实际用途不是很大，因为在做统计分析时有一个很重要的功能就是做同比分析。同比一般是按月进行同比，或者是按日、季进行同比，像这种按任意时间段进行同比是没有任何意义的。

对照研究一下财务报表。我们知道财务报表最小是月报，没有哪个财务报表做的是 2 月 5 日—3 月 4 日的报表。虽然国外的财政年度可以从任

何一个月（比如 4 月）开始，从当年的 4 月 1 日到下一年的 3 月 31 日，但这是一个年度的定义，不是一个任意时间段的定义。为什么财务不会要求有任意时间段的报告，而计算机软件的检索会要任意时间段，唯一的解释就是财务报表已经成熟了，大家已经习惯了规范化的时间范围，而数据统计还不太规范。

通常统计或检索的软件开发人员不知道维度模型，也没有按照年月对数据做特殊处理。把数据按照时间维度显示时，会同时显示每个月的数据，就是同时显示一年中多个月的数据，在这种情况下，任意日就无法使用，除非为每个日期定义一个层次结构，而这是不现实的。

在定义维度时，一个数据表中有很多维度，一般包含度量数据的事实表都对应多个维度，比如在销售订单表中，可能有时间、客户、存货、业务员等维度，所以就像一个事实表周围围绕着多个维度表，画出一个像星星一样的图来，这也是维度模型称为星型模型的原因。

如果一个数据表中有 100 个字段，是不是就有 100 个维度呢？一般会考虑把文字、时间类型字段当成维度，把数值型的当成度量值。不是每个字段都是一个独立维度，而是一个维度有多个层次结构，一个层次结构涉及多个字段。

此外，有些字段数据值不完整，在很多记录中为 NULL 值，不需要建成维度。有些数值比较少的字段，可以转换成度量值而不做维度，如性别字段，男女一般用文字表示，如果做成维度的话维度数量太多，一般就做成一个度量值，改成一个是否男性的度量值——男性值为 1，女性值为 0，这样也方便统计男性和女性的数量。

在星型模式建好后，可以看到许多的维度，每个维度上有不同的层次，如果把它做成图形，就像一个雷达图，从中心放射出许多的维度，每个维度上有很多的节点分别代表层次结构的级别，时间维度的层次结构有年、季、月、日节点。地区维度有省、市、县、区节点，产品维度有大类、小类、产品的节点。

用雷达图看多维数据模型，任何一次查询实际上是每个维度中的一个节点连起来形成的一个多边形。考虑到这种组合非常之多，不可能把所有

组合的数据预先聚合在一起。在开发 OLAP 系统时，科学家研究一些算法，其中之一就是提出一个叫“冰山立方体”的概念，只把最主要的几个组合进行预计算。就像冰山一样，大量的维度组合在冰山的水下面，露出一部分做预计算，或者叫预先聚合。还有个算法叫“外壳立方体”，把立方体外壳部分的维度组合做成预先计算，其他的待查询的时候遇到再做计算。

在 OLAP 系统中，预计算可以设置，从时间和空间两个因素进行平衡。要加快查询速度，减少查询时间，需要牺牲一些空间，多做一些预计算。如果想节约空间，可以忍受一些等待时间，那么可以调到空间最小化。

当维度的组合在预计算的范围之内时，统计速度就非常快，在范围之外统计速度就比较慢。但是，还有一个缓冲设计，就是说你只要查过一次以后，计算结果就被保存下来，速度就比较快。

为了及时更新数据，一般安排在当天下班时间，比如说晚上 8 点到次日凌晨 8 点之间，选一个时间段做增量的数据抽取，把来自数据源中一天增加的数据抽取过来，放到数据仓库的数据库里。OLAP 的服务器从数据仓库服务器里抽取增量数据加在原来的立方体上，按照原来设计的规则对部分数据进行预计算，第二天查询时，就会看到包括昨天数据在内新的统计结果。

不同的软件公司有不同的 OLAP 服务器，如微软公司的 SQL Server Analysis Services（SSAS），Oracle 公司有两个产品，一个叫 OLAP Server，另一个叫 Essbase OLAP Server。

对 OLAP Server 访问目前有两种访问方式：一种访问方式是对 SQL 进行了一些扩充，如 Oracle；还有一种是微软提出的 MDX 语言。MDX 是一个专用的多维数据查询，效率比 SQL 高，在执行时会自动转换成 SQL 语言，比如有一个 MDX 语句会转换 18 个 SQL 语言执行，因此，它是一个适合数据分析的综合性查询。Oracle 里面的 SQL 实际上是模仿了 SQL 的格式，查询命令还会传递给 OLAP Server。OLAP Server 表面上看好像由许多视图组成，实际上这个视图和一般的数据库的视图不一样，是 OLAP 立方体的对外接口。好像访问 OLAP Sevrer 是对视图进行查询，实际上会转换成 OLAP Server 内部的命令，也就是一个查询会在内部转换成很多的

SQL 命令。

通过这两种 OLAP Server 访问语言，可以像一般程序一样进行编程，或者可以在 Excel 里面直接输入 MDX 或 SQL 命令，从 OLAP Server 里面提取数据。这个数据和一般数据库的数据不同，它是一些汇总数据，但比从关系数据库直接统计要快得多。

OLAP Server 实际上有三种设计：一种叫 ROLAP，是关系 OLAP；另一种叫 MOLAP，是多维 OLAP；还有一种叫 HOLAP，是混合 OLAP。区别在于是否把聚合数据放到 OLAP Server 里面去，也就是有没有预计算的功能。

MOLAP 是多维 OLAP，是一种典型的 OLAP，它的做法就是把数据仓库的聚合数据都放到 OLAP Server 里面去，所以速度是比较快的，是现在最常用的 OLAP 类型。

ROLAP 并没有对数据做聚合，还是把数据放在数据仓库服务器中，只是提供了一个 OLAP 的访问接口。如果用 MDX 语句查 ROLAP Server 的话，它的做法是首先把 MDX 语句转换成 SQL 语句，然后分别执行 SQL 语句，在数据仓库所在的关系数据库做计算，把结果整合后反馈。如果关系数据库是普通数据库，ROLAP 的访问速度比较慢，所以 ROLAP 有一段时间基本被抛弃了，直到出现 SAP HANA 这样的内存数据库。

微软的 SSAS 三种 OLAP 都支持，Oracle 的 OLAP Server 根本不支持 ROLAP，开源软 Mondrian 支持 ROLAP。

现在关系数据库新的发展趋势是内存计算，像 SAP 的 Hana。内存计算虽然是关系数据库的架构，但把数据放在内存中，随机访问速度非常快。

内存数据库访问速度快的原因是解决了两个问题。一个问题是硬盘中数据读取速度比较慢，在内存中读取数据要比硬盘快得多；还有一个问题是去掉数据缓存。

在关系数据库里面，为了解决内存数据和硬盘数据之间的读取速度不同，有个缓存命中机制，在读数据的时候会判断数据是在内存中还是在硬盘中，然后再返回数据。如果数据在内存中就会很快，内存中没有就会从硬盘中读。内存比较满，数据很久不用的话会置换到硬盘中。这种转换会

耗费服务器资源。

有了内存数据库以后，ROLAP 可以发挥作用。虽然执行十几个的 SQL，但由于速度很快，效率会很高。ROLAP 带来的最大优点是数据可以实时统计，也就是说把数据从数据源里面提取到数据仓库服务器中，马上就可以查询了，不需要晚上做预处理。当天的数据当天就可以统计，还是比较有吸引力的。当然，这里有一个前提，就是业务系统必须用了内存数据库，但现在很多软件还没用内存数据库。SAP ERP 已经移到 SAP Hana 上，做 OLAP 就可以用到内存数据库。

4.2.4 分主题进行数据分析

在做数据分析中，常常会把数据分析的目标分成不同主题。各主题独立存在。从应用角度来看，一个主题面向独立一个目标，如销售和采购就是两个不同的主题。从分析的数据指标来看，把一批相关指标放在一起作为一个主题进行分析，将其他的指标分为另外的主题。比如，产品销售有产品合同的数据、订单的数据、发货的数据和仓库出库的数据，这些都可作为一组数据放在同一个主题中。而销售订单的数据和采购订单的数据关系不大，就必须放在两个主题中分别进行分析。

主题可以用于对权限的控制，例如销售人员只能看销售的数据分析，因此只有销售分析主题的权限，其他数据放在其他主题中他是看不到的，这样就起到保密和信息安全的作用。

主题的划分也可以从数据的维度角度进行。一般来说，同一个主题中的数据具有相同的维度。例如在销售中，有个重要的维度就是客户，因此跟客户有关的指标一般都应放在销售数据分析主题中。如果指标和供应商有关，那么显然不能放在销售数据分析主题中，而只能放在采购数据分析主题中。

另外，主题的确定还跟维度的多少有关。销售和业务员挂钩，所以有一部分的销售数据会与业务员有关。但是，还有其他与销售相关的数据，比如说应收款。应收款数据只跟客户有关而跟业务员无关，即应收款的数

据比销售的数据要少一个维度。在这种情况下，需要单独建立一个主题，而不是把应收款放在销售数据分析主题中，因为应收款还有一个付款方式的维度，而这个维度跟销售没有关系。这种维度不同且维度个数不一样的数据，很明显不能放在同一个主题中。

当遇到较大主题时，原则上不予拆分。主题的拆分不是根据数据指标的多少来确定的，而是由跟它相关维度来确定。放在一个主题中的数据更便于比较，因为数据只有通过可视化进行比较才能产生价值。若是按照数量拆分，有很多可相比的数据不能放在一个屏幕上，不能相互参照，对决策的意义就小许多。

4.2.5 离不开的时间维度

在数据分析中有一个很重要的维度就是时间维度。时间维度的重要性在于，无论在哪一个数据分析主题中它都是必须具备的维度。换言之，数据分析非常重视历史数据，若只有当前数据而没有历史数据，那么数据分析是不成立的。

时间维度一般可以分为两种层次结构：一为年、季、月、旬、日，这是一个比较完整的时间维度；二为年、周、日。众所周知，周和月不重叠，所以不能放在同一个层次结构中。一般采用是经过简化后的层次结构：年、月、日。

时间维度准确地说是日期维度，实际上只考虑日期，没有考虑时间，如果在工业控制的数据分析中，需要另外建立一个 24 小时的时间维度。

时间维度具有很大的作用，因为很多的指标都与时间有关，有些常用的计算指标，例如同比、环比、年初至今及月初至今，都和时间有关。

时间维度的数据，一般都是合计数：年度合计数、月度合计数和日期合计数。在这里面，日期的合计值加起来等于月度合计值，月度合计值加起来等于年度合计值。可以通过数据向下钻取，从任意一年看到任意一月，再看到任意一天合计数。反过来，通过上卷（向上钻取）也可以从日看到月，再看到年的合计数。

一个概要页面中，会有一个年度至今的数据指标。因为做一个仪表板，常常要看看当年年初到统计时为止的合计销售额，即从当年 1 月 1 日开始到当天的合计值。

中国的会计年度自公历 1 月 1 日起至 12 月 31 日止，但有的国家是从每年的 6 月 1 日到第二年的 5 月 31 日，以及从每年的 9 月 1 日到第二年的 8 月 31 日等。因此，时间维度又多了一种定义，即财政时间维度，相对的日历时间维度。在数据分析时，需要分别对这两个维度进行分析。

时间维度表示的是历史的趋势，一般用折线图表示。利润、收入这些可以合计的数据指标可以采用面积图。

在一般的软件开发中，为了既简化设计，又能满足任意需求，在查询或报表中，只要涉及时间条件的，一般采用可以同时输入开始时间和结束时间。因而，很多人在数据分析中也希望有这样的功能。

在事务软件开发中，由于不需要考虑层次结构，所以输入任意开始时间和结束时间是一种比较省力的实现方式，实际上是对时间的查询需求没有细化。而在数据分析处理中，一般不支持这种处理。数据分析建模中，需要按照预先定义的维度进行预处理。随意输入的开始时间和结束时间无法进行预处理。由于分析需要进行同比、环比计算，任意时间段的处理也无法计算同比环比，即使有也没有任何意义，因为一般都讲 5 月环比增长，不会讲 5 月 10 日到 5 月 21 日的环比。

随着数据分析的深入应用，人们也会逐渐接受按照固定的时间进行分析的习惯。实际上，在微软开发的多维查询语言 MDX 中，如果不考虑数据查询的时间，或者在数据集比较小的情况，它的任意时间段查询是可以实现的，但计算查询相关的同比、环比分析就会比较困难，且不利于大家共享分析结果。

4.2.6 通过时间分析数据

时间维度是数据分析中最主要的一个维度，每一个数据都应该有时间维度。时间维度表示的是数据的历史信息，历史信息里保存了非常多可以

分析的迹象，从中可以发现很多的问题。

一般来说，在时间维度都是用折线图来表示，因为折线图表示的是连续的意思，在一定的时间内变化不大，而且有一定的关系，所以用折线图把不同时间点的数据连起来可以看到变化趋势。还有一些值可以用面积图来表示。面积图可以表示出这段时间之内的总数，比如说销售额，如果用一年 12 个月销售额的面积加起来应该就是全年的销售额。有些数据之和是没有意义的，比如说库存的余额，它不能相加，只能用折线图。而且有的时候为了方便比较，会把很多不同的指标放在一起，如果用面积图就会互相遮挡，不利于互相比较，所以也会用折线图。

从表明时间趋势的折线图上，无论是不同年份的比较还是同年不同月份的比较，如果不对数据钻取，是看不出太多有价值的信息的。在折线图上看到某一年或者某一个月中的相比数值特别大或者特别小，可以钻取下去，看看到底是由什么数据引起的。某一个数据如果在年度数据中数值增加比较多，可能是平均增加得比较多，但更多的情况只是某一个时间区间增加得比较多。这样的话，你就可以在大数据集中找到异常的数据。

比如说，在一个数据分析主题中汇集了 10 年的数据，每天新增 1 000 条数据，10 年的数据显然会是很大的量。如果其中有一天的数据甚至是有一条记录有异常，比如说一天销售特别多，在分析的时候就可以通过钻取来找到这条特殊的记录。具体的发现方法是这样的：由于这条记录会导致该年的平均值比较大，所以先从日平均销售额的时间趋势看，会发现其中一年的平均值比较大，钻取这年数据，看到按月分布的平均值，又发现其中有一个月的平均值比较大，再钻取到该月数据，找到平均值大的日期，然后再看销售记录明细，就能马上找到这条数据，看到明细状况。

单纯看时间维度，有时难以看出特殊的问题，这时可以从其他维度看到一些异常，再结合时间维度看历史情况，就能找出问题所在。比如发现有一个客户的销售额增长特别快，我们会在客户维度上锁定这个客户，看这个客户的历史，从他的历史看他是新客户还是老客户。结果发现他虽然是一个老客户，但原来采购比较少，最近采购特别多，可以安排销售人员了解一下他最近采购多的原因，有针对性地提供服务。

另外，时间维度要关注是同比和环比。同比是看每个月的数据和上年同月数据的比较，环比就是这个月数据和上个月数据的比较。如果是日期的话，同比是指这天和上年同一天的数据比较，环比是和上个月同一天的比较。

同比可以按照某一个维度，比如客户，来找出增长最大的客户。具体就是对同比进行排名，找出同比数值最高的 10 个客户。为了避免有些合计数值比较小，但同比变化比较大的情况，可以对金额和同比同时进行排序，也就是只对数值排名在前 70% 的客户进行同比排序。所以，虽然合计数值的同比也可以在历史时间维度上比较，但要直接找出一些特殊的对象，这种同比 TOP10 排名揭示的信息还是非常多的。

时间概念里，还有一个叫年初至今或者月初至今。年初至今就是指从本年的 1 月 1 日开始累计的数据，用于在概要上看到某一个指标累计的执行情况，可以看到到目前为止的业绩。如果做一个年初至今数据的同比，比较上年同一天的累计业绩，可以在日常数据的监控中，发现经营的问题。如果发现数据同比下降了，就要去找出下降的问题所在。月初至今是指从本月 1 日开始的累计数据。

4.2.7 空间维度直观地显示数据

在数据分析中常常会遇到空间维度，空间维度最典型的情景即是行政区域。

空间维度可能是一个独立的维度，即维度的层次结构中每个级别都为空间，比如按区域显示销售额或 GDP。

空间维度也可能是维度的一部分，即层次结构中上层几个级别为空间，下面的级别与空间无关，比如销售分析中，客户维度所在的区域为空间，客户名称与空间无关。

空间维度虽然可以用文字描述，但若用地图，将会非常直观。因此在很多数据分析应用中，常常会把一些描述区域的维度或级别用地图来显示。

人们通过在地图上标识不同的颜色，来表明数据的大小。点击地图不

同的区域，可以进行钻取，比如从全国地图上点击省份可以转到地区，单击地区可以转到县城、县城可以转到街道。

4.2.8　数据的可视化钻取

利用维度模型建模，可以实现多维度并行操作，通过数据钻取可达最小粒度数据，即从合计数据钻透到明细数据。虽然最小粒度数据是一个，但可以从多个维度分析，从多个维度可从同一个数据集中钻取到一个子集，见图 4-4。

数据的粒度越小，可以分析的维度越多。每次合计，都是以损失一个或多个维度（或层级级别）为代价的。比如说最小粒度数据以日记录，如果生成月度数据，则无法分析数据按日的分布情况，也无法按周分析每周的数据变化。



	单据日期	订单号	组织	地区	客户分类	客户名称	存货	部门	业务员	销售数量	销售金额	销售收入
1	2011-01-16	SO1101160017	米奇宝（北京）	东北		沈阳康贝商贸有限公	金装幼儿配方奶粉450g	销售部	张霞	22888.00	3015494.00	2577345.30
2	2011-02-17	SO1101170023	米奇宝（北京）	华东		扬州金冠臣商超	婴儿营养配方乳粉250g	销售部	张霞	7900.00	687774.00	587841.03
3	2011-02-17	SO1101170022	米奇宝（北京）	东北		沈阳金冠臣商超	婴儿营养配方乳粉250g	销售部	张霞	1980.00	191664.00	163815.38
4	2011-02-16	SO1101160015	米奇宝（北京）	东北		沈阳康贝商贸有限公	米奇宝新生儿配方奶粉900g	销售部	张霞	1200.00	235980.00	201692.31
5	2011-02-24	SO1101240027	米奇宝（北京）	华东		苏州鑫义贸易有限公	米奇宝婴幼儿配方奶粉900g	销售部	张霞	1960.00	132251.00	113035.06
6	2011-02-17	SO1101170025	米奇宝（北京）	华东		苏州鑫义贸易有限公	婴儿营养配方乳粉250g	销售部	张霞	620.00	53977.20	46134.36
7	2011-02-17	SO1101170024	米奇宝（北京）	华北		烟台市丽宝贝食品有	婴儿营养配方乳粉250g	销售部	张霞	850.00	76228.00	65152.14
8	2011-02-17	SO1101170021	米奇宝（北京）	华东		苏州鑫义贸易有限公	速食营养奶糕300g	销售部	张霞	1000.00	67475.00	57670.95

图 4-4　多维度数据钻取到最小粒度记录

通过交互操作实现可视化的数据钻取，而不是在数据上选择，可以大幅提高数据分析的效率和价值。以下为在日期维度从年（见图 4-5）到月（见图 4-6），再到日的钻取（见图 4-7）。

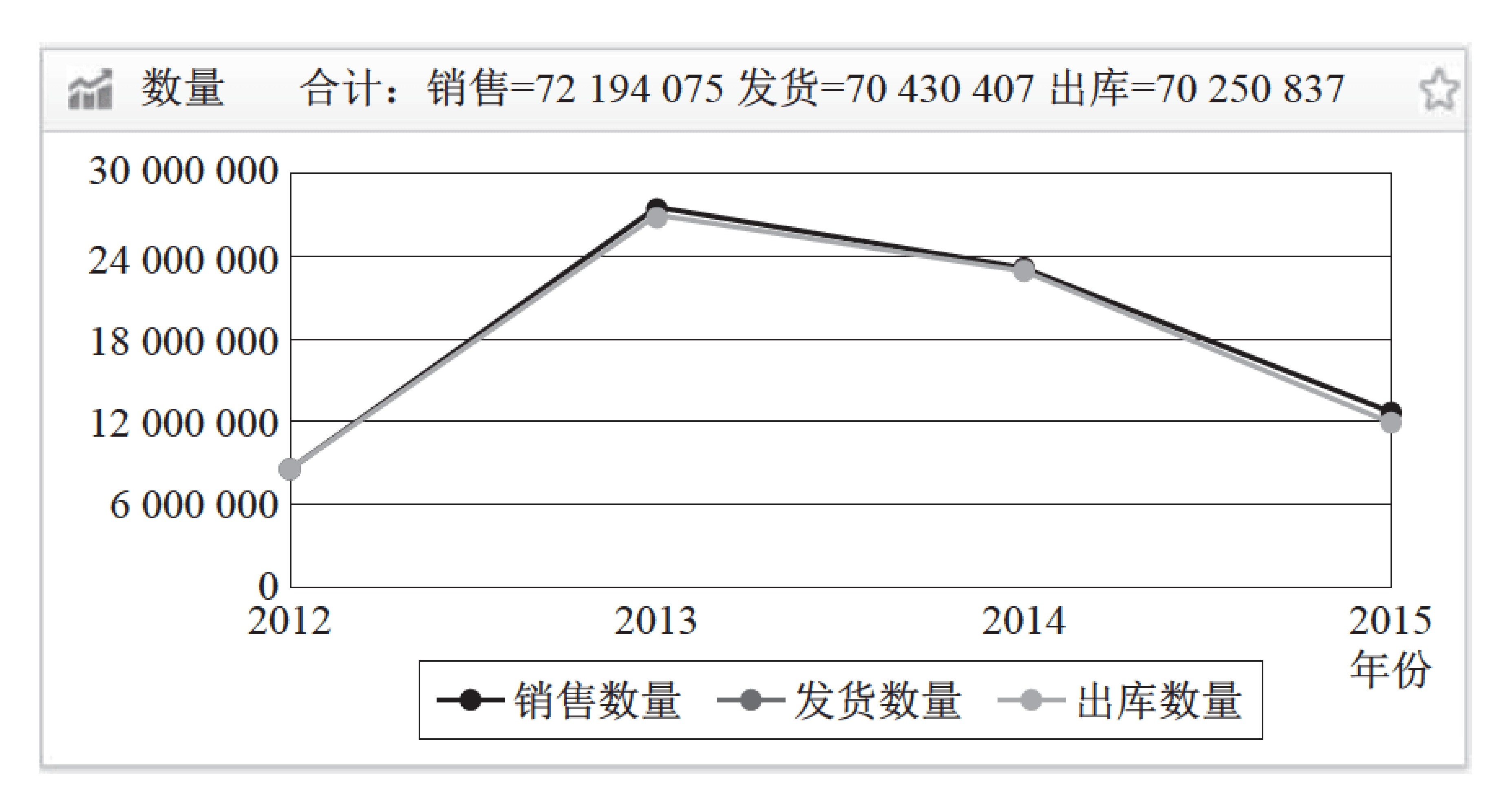


图 4-5 按时间维度的数据钻取，2012—2015 年年度数据

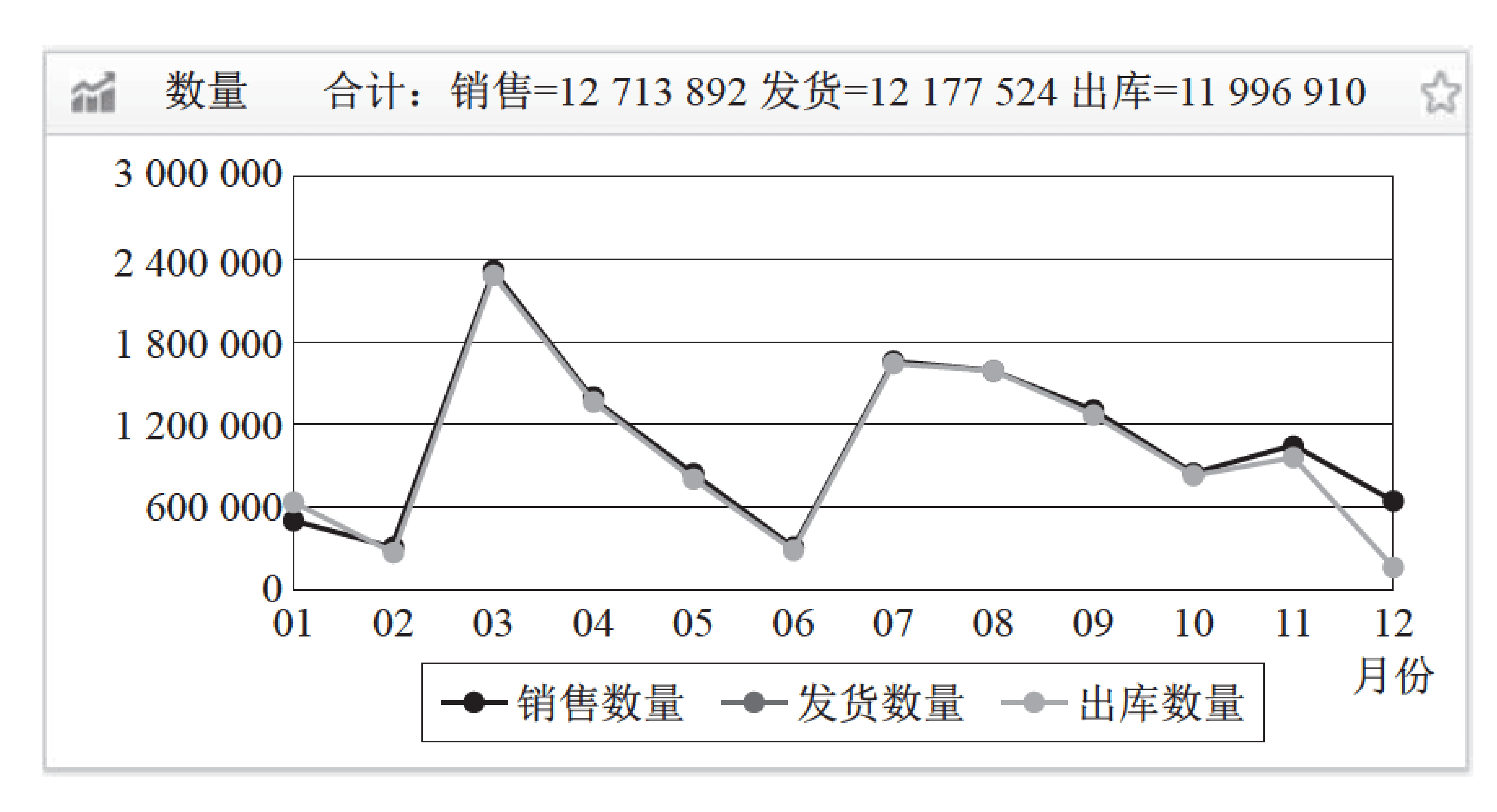


图 4-6 按时间维度的数据钻取，2015 年月份数据

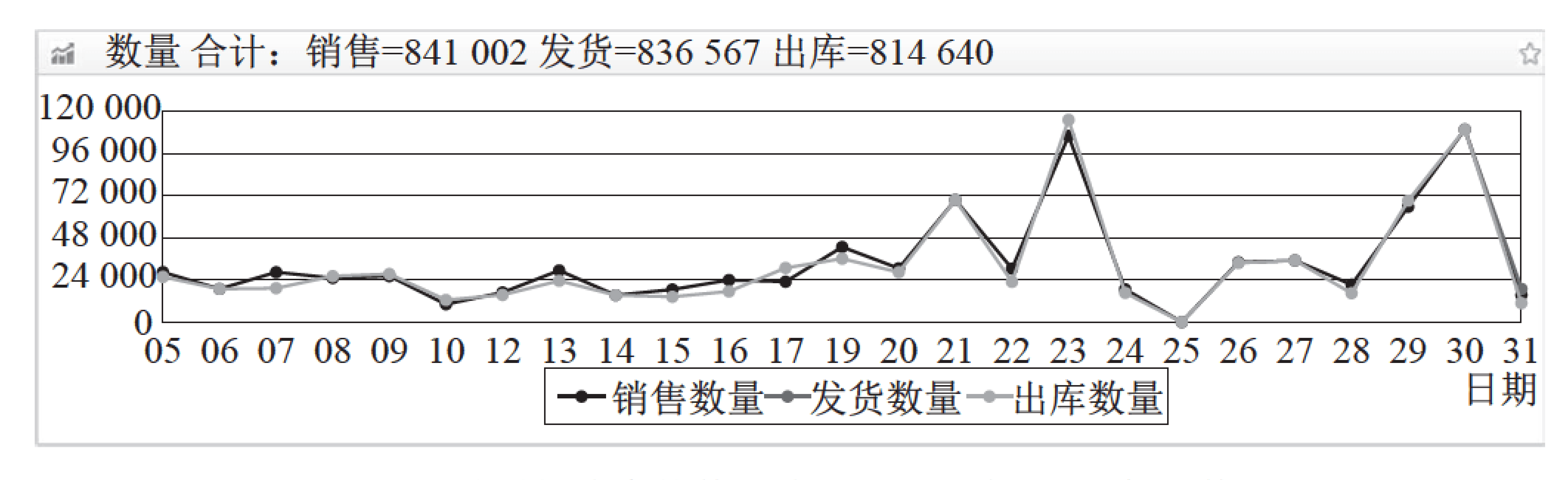


图 4-7 按时间维度的数据钻取，2015 年 10 月每日数据

图 4-8、图 4-9 是从客户维度向下钻取，从一个省到该省的客户：

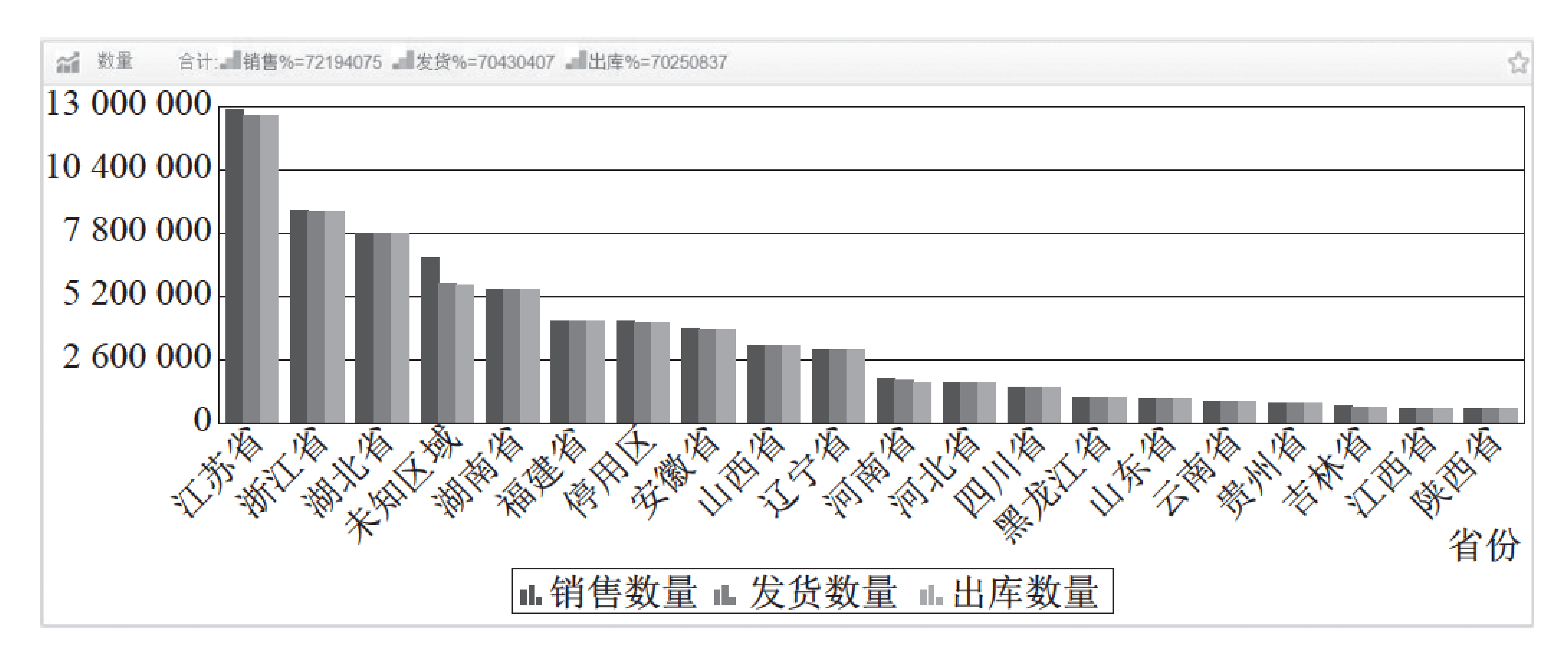


图 4-8　按客户维度的数据钻取，全部数据

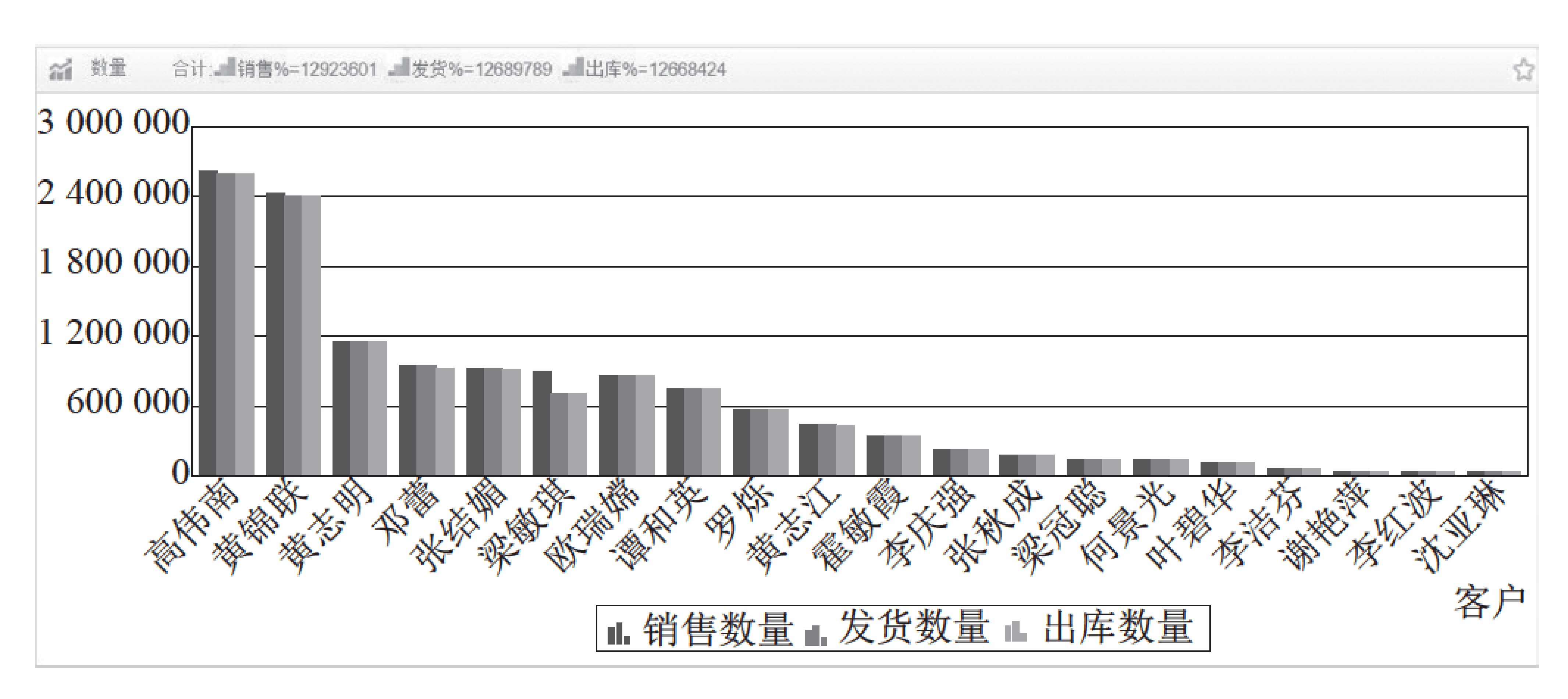


图 4-9　按客户维度的数据钻取，江苏省数据

4.2.9　用OLAP提升统计速度

数据时代的主要特征就是处理的都是大数据，大数据处理如何提高速度非常重要。

一般的关系数据库，现在又称作 OLTP 数据库，它的处理花费时间会随着数据量的增加指数增加。当然这里的数据处理指的是求合计或平均值，而不是一个索引查询。因为索引可以快速定位到指定数据，花费时间不会指数增加。

数据分析更多的要涉及全体数据，所以索引技术是无效的。在这种情况下如果用 OLTP 数据库来处理类似的问题，它查询的时间会迅速地增

加，这个增加就像一个指数曲线，所以根本无法使用，因此，需要专用的OLAP数据库。

OLAP数据库的特点是，如果查询很小的数据时，它的速度也不会很快，相对OLTP来说可能还慢，但如果数据很大，速度也不会很慢，可能会稍微慢点，但随时间增加比较平缓。用Hadoop来实现类似的分析时也会有类似的问题，就是即使对一个非常小的数据集做合计，时间也需要几十秒，但如果数据量非常大，速度却不会很慢。

在大数据的应用中也会用到关系数据库，一般的关系数据库用作数据仓库数据库，这种数据按照一个数据仓库模型来组织数据，但数据仍然放在关系数据库里。有一些人认为建了一个模型以后就是完成了建模，实际上这只是模型的一个图纸，并没有发挥实际效益，必须把数据整体地装到OLAP Server里才行。OLAP Server使用首先要进行配置，配置的过程就是和数据仓库数据库里的数据表建立起一个关系，在具体构建OLAP立方时它会从这里面读取数据。

OLAP立方的构建不是简单地把数据拷贝到OLAP Server，而是在里面建立一个内部结构，会对数据作一个特殊的处理，主要的处理就是进行预计算，也就是说把可能用的一些条件的组合结果预先计算好存在这里。如果查询正好是这些条件组合的话，就能马上反馈汇总的结果，不需要再进行计算。但如果查询没有命中这种组合，就必须进行计算，这时查询速度就会比较慢。每次计算的结果会存起来，如果下次有相同的组合时还可以使用，而且速度会比较快。

预计算保证OLAP Server使用越多，速度越快。至于这个做多少预计算合适，这里面有一个平衡。如果预计算太多，占用的空间会比较大，而且初始化处理的时间也会比较长，但很多预计算的结果不一定用得到。如果预计算太少，虽然占的空间比较少，但计算时速度会比较慢，因为条件稍微变一下就会超过预计算的范围。在Oraccle OLAP里面，预计算的缺省比例是40%，但也有一些书上介绍一般15%就可以了。

如果想把这些条件组合都做预计算，这在某些书中有一个专有名词是“维灾难”，也就是说维的组合是一个非常巨大的数据，是不可能实现的。

现在也有一些不用 OLAP Server 的软件，自己会做一些轻度汇总。但这些轻度汇总和它的功能相关，就是说功能中需要哪些数据，才会对这些数据做汇总，那这样就明显将它与需求挂钩了，无法满足不确定的需求。

采用 OLAP Server 的主要目的是做一个无方向的数据挖掘，也就是说需要面向任意需求。数据汇总的预计算不能认为确定，而要有算法模型的支撑。不同公司提供的 OLAP Server，其 OLAP 立方的预计算是要依赖一些算法的，比如说冰山立方体算法、外壳立方体算法等，它通过一些算法相对科学地预先计算一些合计结果，保证查询的命中率尽可能得的。

在鹰眼技术里，还在程序这一层做了一个缓存。虽然 OLAP Server 的查询速度已经够快了，但因为在实际应用中常常需要从多个数据表中提取数据，而且条件的组合也比较复杂，查询需要等待一段时间。如果客户在等待一个新的数据分析图形，这个时间是可以忍受的，但如果刚刚看过一个维度的数据，在看下一个维度之后又返回，就像网页的“上一页”一样，等待这个时间会感觉太长，因为客户概念中这个页面已经做了缓存，应该很快出现。因此，在程序里又做一个缓存，让客户感觉页面已经缓存在本机了，老页面点回去就非常快。这种缓存是基于 OLAP Server 基础上的二次缓存。

4.2.10 数据可视化加快对数据的认知

关于人类视觉的重要性，大卫•麦克坎德雷斯在他的 TED 演讲《数据视觉化之美》[9] 中说道，视觉是人类五大感观中处理信息速度最快、数量最大的。他引用丹麦物理学家 Tor Norretrandersde 的研究成果，意思是，人类大脑作为高级信息处理器官，有很多的信息来源：通过眼睛输入的视觉信息、通过耳朵输入的声音信息、通过鼻子输入的气味信息、通过舌尖接触的味觉信息、通过皮肤触摸接触的触觉信息，最终，这些信息都会传输到大脑中。人脑对这些信息接收速度区别非常大，就像计算机联网的带宽一样，不同信息来源带宽不同。带宽就是同一个时间段所传输的字节数，带宽宽的话下载一部电影可能非常快，带宽窄的话可能非常慢。因此，不

同的带宽在同一时间段从网络上获取的信息量是不一样的，如图 4-10 所示用面积大小标明不同感官的带宽。



图 4-10　不同感官的带宽

同理，我们大脑在同一段时间对不同的信息渠道接收的信息量也不一样，信息量最大的就是视觉。我们通过眼睛获取的信息要远远超过其他方式，其次是听觉、味觉，最少的是触觉。

假如我们要听一段五分钟的文字汇报，将文字要传递的信息绘制成一张图表的话，我们可能只需要花 1 秒的时间就能将所要了解的信息传递到大脑。

简而言之，如果把信息用图形的方式展示，再输入大脑的话，效率是最高的。这就是现代大多数人更加愿意利用图形来展示信息的一个重要的原因。

数据可视化还能避免对数据的误读。学者埃姆雷•索伊尔（Emre Soyer）和罗宾•霍格尔斯（Robin Hogarth）做了一个有趣的研究。研究对象是对数据毫不陌生的经济学家，三组经济学家分别回答了一组数据的同样一个问题：一组经济学家拿到的是数据和数据的标准统计分析，72% 的人给出错误答案；另一组拿到的是数据、统计分析，以及一张图表，答案错误率仍高达 61%；还有一组只拿到图表，仅有 3% 的人回答错误。[10]

从另一方面来看，因为数据表示的大多是“1、2、3、4”这种数字信息，我们知道数字是人类的一种发明，尤其是阿拉伯数字，是人类发展史

上一项伟大发明，而阿拉伯数字中的零是人类发展史上更加伟大的一个发明。因此，数字是人类发明出来的，而不是与生俱来的。人类对数字的判断、识别是后天训练的结果，而不是天生的。但是，人类对这种将大与小、多与少表示成图形的识别就是天生的能力。对信息的识别，以图形展示的方式相对其他方式而言，人的负担会减少许多，这也是把数据可视化的原因。

4.2.11 用内存数据库实现实时数据分析

虽然大多数情况下实时数据分析并不太需要，但不乏有些人有这方面的需求。

数据分析一般分为以下几个过程：

（1）数据的初始化。初始化把历史上多年的数据一次性抽取到数据仓库中来，这需要花费较长时间。

（2）增量抽取。把最近一天新增加的数据抽取到系统中来，花费时间较少。增量抽取工作一般安排在半夜进行，当业务系统停用以后，在系统资源比较空闲的情况下去提取它的数据，这样的话对业务系统的使用没有任何影响。

（3）OLAP Cube 构建。把数据放在 OLAP Server 中，并进行预计算。

这样处理后，第二天就能看到头天及以前的统计数据。

如果要实现实时数据分析，就必须随时进行数据增量抽取和 OLAP 构建，比如设计一个定时程序，每 1 分钟或者 5 分钟去读取一下业务部门的系统。这样的话，很可能对业务系统的运行产生影响。另外，OLAP 的处理也需要时间，不能在 1 分钟或 5 分钟能完成。如果对数据实现 1 小时延时的数据分析应该可以做到的。

现在技术上也在向实时数据分析方向发展。SAP 公司的 HANA 系统支持实时数据分析。它的原理是实现一个内存的数据库，所以汇总的速度非常快。因此，不再需要构建 OLAP Cube。

当然，这里有几个条件：

（1）业务系统必须在 SAP 的 HANA 中运行，即它的业务系统（OLTP

系统）必须使用 HANA 数据库，而不需要把 OLTP 数据读到 HANA 中，保证数据读取在内存中进行。

（2）需要 ROLAP 的支持，用 ROLAP（即关系 OLAP）这个模型去直接读取关系数据库中的数据。

HANA 提供了建模工具 Modler，在建模后可用 MDX 语言访问 Cube。目前主要问题是除了 SAP 软件以外，很多常用软件并没有移植到 HANA 上。

总体而言，未来的趋势是数据库全部变成内存数据库，OLAP 以 ROLAP 为主而无须 MOLAP。

由于用于数据分析的 OLAP 数据库的格式与事务处理系统的 OLTP 数据库不同，所以即使有了内存数据库，定时加载步骤还是不可缺少，除非只是直接从 OLTP 数据库中读取几个指标值，否则完全的实时数据分析还是需要耗费很多资源，具体就看是否值得这样做了。

4.3　改变思路

4.3.1　建立基于真实数据的KPI

数据时代，关注和处理的应该是反映真实世界的数据，而不是人造的数据。

管理发展的历史上，KPI 是一个重要的管理工具。企业为了更好地管理员工，设立了一系列的关键绩效指标（Key Performance Indicator，KPI）。KPI 作为一种企业绩效管理工具，把企业的战略目标分解为可操作的工作目标，可以使部门主管明确部门的主要责任，并以此为基础，明确部门人员的业绩衡量指标。建立明确的切实可行的 KPI 体系，是做好绩效管理的关键。

KPI 制度是人为设定的，它的标准完全是人为的，而且跟效益、收入

密切相关，所以导致企业进入两个误区：一个误区就是公司花很多的精力做 KPI 的统计和分析、打分，有很多的管理人员花费很多时间参与其中；第二个误区是员工靠 KPI 的引导工作，KPI 考核的就做，不考核的就不做，而不是以客户的利益和公司股东的利益为导向。

基于数据的管理，即所谓量化管理，是一个正确的方向。但关键是如何得到量化管理所需要的数据。数据来源有两个：真实数据和人造数据。真实数据就是在企业经营中由事务处理系统产生的数据，这要求企业信息化水平较高，员工的许多日常工作都通过信息系统完成，并留下记录。人造数据就是通过同事及领导打分得到的数据，这些数据可能是有些行为无法在信息系统中记录，更有可能是一个企业信息化水平很差，没有记录实际经营情况的数据。

现在企业大多数采用混合模式，即从信息系统中获取数据和人造数据结合。如果刻意追求 KPI 体系的完整性，不以结果为导向，希望员工的所有工作都可以量化追踪，就难免出现很多人造数据。鉴于 KPI 制度的弊端，现在有很多的公司，比如发明 KPI 的埃森哲公司已经抛弃了 KPI 模式。

一个理想的 KPI 制度，应该利用现有的数据，根据信息系统中可以获取的数据编制，而且 KPI 仅作为一个收入和晋升的参考，而不是直接挂钩。当然，为避免人造数据，需要加强企业信息化建设，更要加强对数据的开发，比如建立企业级数据仓库。

4.3.2　为实现工业4.0建立数据基础设施

工业 4.0 作为制造业未来发展的一个目标，是正确的发展方向。工业 4.0 的核心是实现 CPS（Cyber-Physical Systems）系统，也就是虚拟实体系统。

什么是 CPS 系统呢？

现在存在两个世界：一个是实体世界社会，即传统的制造业，也就是物质制造；第二个是虚拟世界，即我们的信息世界或虚拟世界。从物联网和人联网的角度来说，物联网涉及实体世界，关乎物到物的沟通和控制；人联网涉及虚拟世界，关乎人与人之间沟通和控制。

工业 4.0 的目标就是实体世界和虚拟世界的融合，物联网和人联网的融合。比如在一个社交网络上可以对产品提出需求，对设计进行修改，然后这些信息可以传输到以物联网为基础的这种智能制造系统上，智能制造系统就能根据这种信息制造出一个实际的物理产品，并且通过物流发送给需求者。

某人过生日，我们可以有两个祝福的方案：一个是实体世界方案，另一个是虚拟世界方案。在实体世界过生日，去蛋糕店买个蛋糕，写上名字、年龄，过几天去取或者送货上门。在虚拟世界过生日，可以在 QQ 上送生日礼物，比如送一个虚拟的蛋糕，通过 QQ 发送。在实现工业 4.0 以后，这两个世界可以完全融合在一起，在 QQ 上看到别人生日就可以送个蛋糕，选中一个蛋糕以后，这个信息会传送到实体蛋糕店，蛋糕店根据这个信息做出蛋糕、打上名字和年龄，再通过物流送到指定的对象家里去。

大家知道，现在即使送生日蛋糕简单的虚拟实体系统也没能实现。如果是复杂应用呢？比如说我们想定制一辆汽车，如何网络定制、下单，实体制造、送货，就更加复杂，需要对整个制造体系进行改造。

仔细分析虚拟实体系统的实现流程，可以分解为几个共同具备的过程：第一，通过人联网收集客户的需求；第二，通过物联网了解现在设备的运营情况；第三，下达具体生产的指令；第四，得到生产状态的反馈。

客户既然是定制的，就需要跟踪。比如，一辆汽车的生产周期要 30 天，显然对于客户来说，他在虚拟平台上面下了购买指令后，肯定需要跟踪生产的状况，确切知道什么时候能够提货。

综上所述，实现工业 4.0 的前提是已经有了比较成熟的数据技术。必须了解设备的运营情况，知道现在有多少台设备，设备的生产能力如何，生产状态如何。要通过物联网技术采集这些数据，而且让管理者能看到这些数据，这些数据最终还应该能够直接通过人联网送达客户。

比如说现在要订一个蛋糕，显然先要知道周边方圆几公里之内有几家蛋糕店，了解蛋糕店的订单情况和执行进度。如果要求蛋糕在下午 1 点钟送达，这种情况下，先要查一下附近哪个店现在具有按时送达能力。这个能力不是根据口头承诺，而是根据数据判断出来的。比如，做一个蛋糕需

要两个小时，那么就必须在 11 点之前开始制作，如果有蛋糕店 12 点钟才能开始制作就来不及。可以根据设备数据判断制作开始时间，蛋糕店有多少台设备，设备的生产能力，设备现在的工作负荷情况，设备的检修情况，今天总共有几台设备在正常运行，据此判断这个单子能不能排队 11 点钟前开始生产。下单以后，还需要了解制作情况的反馈：制作过程中是否按计划执行，中间有没有出现故障。

所以，实现工业 4.0 首先要求实现智能制造。在现有设备基础上第一个就是设备必须智能化，必须能采集数据；第二个数据是能够联网的；第三个数据是能被共享的。如果没有这些数据，那就根本不可能实现物联网的功能。

另外最大的问题是关于人联网的问题。显然要实现工业 4.0，人联网非常重要。虽然现在互联网很发达，单独实现人联网不是问题。但如果需要一个企业独立实现工业 4.0 的话，既要求该企业具备先进的制造能力，又具备互联网的开发经营能力，软硬都要通，这是不太现实的，也违背现在社会分工细化的要求。如果有其他社会组织能提供全部或部分人联网的功能，让制造业企业专注于制造，则实现工业 4.0 更为容易。

因此，数据的社会化的运营，对工业 4.0 是一个非常重要甚至是一个必要的条件。我们必须通过对数据领域的规划和布局，使得社会的信息，就是人联网的信息能够非常容易地获取和传送。而这种社会化的运营不是靠像中国的阿里巴巴或者腾讯那样的第三方企业来实现，而是建立一个基础设施，就像公路和铁路一样，然后在基础设施上再连接这些相关的企业。

我们知道，传统行业设施建设比较成熟，新兴产业完全可以借鉴传统产业的经验。比如，交通领域的机场是公用设施，公路、铁路也是公用设施，在上面跑的飞机、汽车、火车可以是不同个人或公司的。中国铁路总公司这种企业，既造铁路，又造火车，还运营线路，只能是国有企业的运营方式，不可能由私营企业来实现。但现在中国铁路也在探索吸收社会资本的加入。所以工业 4.0 的虚拟世界建设也需要分为公用设施建设和企业独立建设两个层面，数据的公用设施建设对工业 4.0 的实现非常重要。

4.3.3　主动抽取数据实现数据集中

加强顶层设计和统筹协调，大力推动政府信息系统和公共数据互联开放共享，加快政府信息平台整合，消除信息孤岛，推进数据资源向社会开放，增强政府公信力，引导社会发展，服务公众企业。

国务院《促进大数据发展行动纲要》

在数据时代，如何汇集异构系统的数据，解决数据孤岛，是一个技术难题。

目前，一般要求从下到上报送数据。就是先制定一个数据标准，然后要求数据源的企业开发专门软件，将数据发送到上级数据中心。这种数据集中法一般要通过行政命令来执行，数据源单位的工作量比较大。

数据报送的方式对上级单位来说，是集中数据的一种比较简便的方法，只需要制定数据交换标准，要求下属单位按标准报送数据即可。自己工作量比较小，而下属单位工作量比较大。如果数据集中的任务不能按时完成的话，主要责任在下属单位，所以比较适合具有行政隶属关系的单位。

数据报送软件的开发有比较大的工作量。由于不同软件由不同软件公司在不同时间开发，在运行一段时间后，要找到原开发公司和原开发人员比较困难，也难以面向社会重新招标，所以开发成本会比较高。

由于数据分析技术不够成熟，数据源单位无法知道发送出的数据对自己有什么用处，分享不到数据分析的具体效果，因此也对数据集中缺乏积极性。

综上所述，用数据报送方式解决数据集中问题，造成的后果可能是：成本高、拖延时间长、风险大、效益低。

解决以上问题的一个最好方法是“取数据”，就是由数据中心去各个数据源单位定时抽取数据。用取数据代替送数据，可以将分摊到多个软件开发公司的工作统一委托给一家开发公司。数据源企业只需开放数据，减轻了数据集中的阻力。

对数据分析的结果应该及时分享到数据源单位，这有利于促进数据源单位配合工作的积极性。

假设在医疗数据的收集中，不需要医院投入太多，就能分享一个地区

或者全国的医疗收费数据，那么在医院医疗资源的配置和收费标准的制订上，都会对医院有非常大的帮助，所以各医院会积极配合这件工作。

4.3.4 统计数据从报送到抽取

政府统计数据往往出现很多问题：相互标准不统一、信息共享不畅、数出多门及数字打架，导致数据的不一致。从填报企业来说，面向财税、审计、统计等不同的部门往往从各自利益角度出发填报不同的数据及几套报表的现象严重影响了统计数据质量。

若按照传统方式收集统计数据，这种问题难以避免。解决问题的方法是改报数据为取数据。报数据的话要经过人手，这中间会有动机和机会对数据进行修改。取数据，因为数据源只有一个，不经人手，没有机会修改数据，数据不需人工汇总而是自动汇总。

保证数据准确的另一个方法是数据产生和统计目标分离，也就是数据的产生和保存应该是出于其他的目的，该目的和统计不相干。比如取财务数据，凭证的数量比较多又有复杂的钩稽关系，如果要作假工作量比较大，并且财务软件的使用并不是完全为了统计的需要，所以基于财务凭证得到的统计数据相对准确性比较高。如果只看财务数据的生成报表，虽然有一定的钩稽关系，但由于报表相对数据较少，并且不太复杂，就比较容易作假。

类似地，如果根据出库数据来统计企业销售收入，得到的结果就会比较真实。因为企业记录进出库数据主要用于计算库存，用于内部管理的目的，而且数据量比较大，难以作假。当然这种数据取法的工作量比较大，是一个长期的、艰巨的工作，但也只有按照这个目标去做，才能彻底解决这样的问题。

4.3.5 改进数据分析工作流程

现在很多的互联网公司以及游戏公司对大数据的应用非常重视，很多已经在 Hadoop 技术上进行了投资，但大部分可能只是设置了数据分析

的岗位。

现在来分析一些大数据分析工作的流程，结合自助分析系统的使用，提出改进意见。

互联网公司拥有大量的数据，这些数据可能存在普通的关系数据库，比如 Oracle，或者存在大数据的分布式文件系统，比如 Hadoop 中。

在需要使用数据的时候，就从这些数据库中提取数据。因为 Hadoop 可以利用 Hive 工具用类似 SQL 语言读取数据，所以我们把 Hadoop 加 Hive 也当作普通的数据库来对待。取得数据之后，用一些统计分析工具来分析。常用的统计分析工具有 SPSS 和 SAS，也有简单的就用 Excel 和 Tableau 等可视化的工具进行分析，最后把分析结果和生成的统计图形写在 Word 报告或 PPT 演示中提交给领导或业务部门。软件开发人员和业务运营人员根据这些数据分析的结果修改程序或者调整业务，比如电子商务公司会根据销量数据分析来调整货物的存货。

在以上数据利用过程中，关键在于对数据的访问必须有一个长的流程。一般来说，数据分析人员对 IT 系统的数据库不是很熟悉，他需要给 IT 人员提出需求，由 IT 人员协助获取数据。数据可能已经存在数据库中，只需直接写一个 SQL 语言命令就可以提取。也有可能数据比较复杂，存在多个数据表中，需要通过连接来提取。还有一些汇总数据，比如说按区域的销售数据区汇总，可能需要经过比较长时间的汇总计算才能得到。更有一些数据，可能在数据库中没有保存，需要修改程序，扩充功能，增加需要采集数据的字段，才能获取数据分析人员需要的数据。所以，数据分析人员需要等待得到数据，而 IT 人员可能业务比较繁忙，向他索取数据的人比较多，或者由于开发水平的限制，不能及时提供给数据分析人员，所以数据分析人员的工作受制于 IT 人员。

同样，业务人员为了更好地开展业务，则需要来自数据分析人员的分析结果。也就是说，他需要向数据分析人员提出请求，然后由数据分析人员通过分析，对业务的工作提出指导意见。那么，数据分析人员由于分析工作繁多或者是人手不足，甚至是 IT 部门人员的配合有问题，也难以响应业务部门人员的请求或者可能不能及时响应。

图 4-11 显示的是现在数据分析工作的流程。

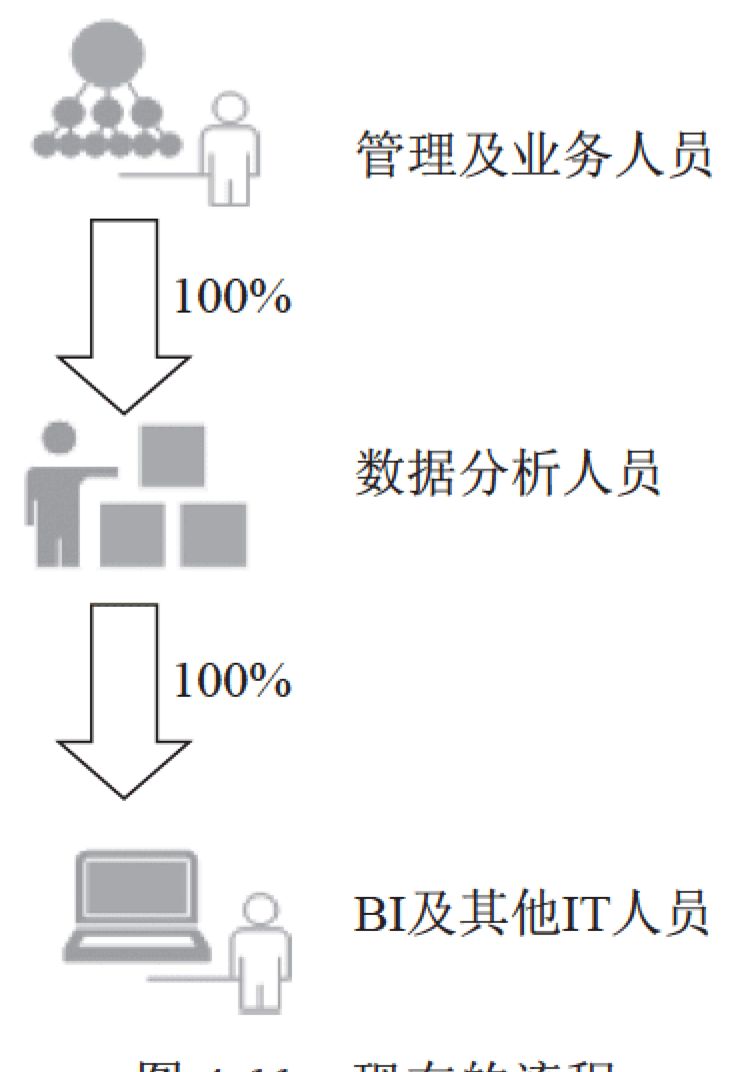


图 4-11　现在的流程

由此看来，要提高组织的运行效率，对数据分析工作改进的最好方式就是业务部门人员可以跳过数据分析人员和 IT 部门人员，直接从数据中发现运营中的问题。

如果我们有一个自助分析系统，可以直接从数据源抽取数据并且生成统计图形，业务人员可以直接通过图形发现问题，就可以大大提高企业的工作效率。

当然，有了自助分析系统以后，IT 人员和数据分析人员并不是无事可做。因为这种分析虽然采用了过度设计，但肯定不能百分百地满足数据分析需求，比如说本身在事务处理系统里没有数据，这肯定无法进行分析，所以还是需要 IT 人员去增加数据的采集内容。同样地，有些复杂的数据分析还是需要数据分析人员进行分析。

所以，自助分析系统可能只能满足 80% 的通用数据分析需求。原来数据分析人员全部利用 IT 人员来获取数据，有了自助分析系统后，现在有 80% 可以通过系统来获取数据。如果需要进一步分析，可以把数据导入平面文件中，通过其他的软件，比如 R 语言，来进行数据挖掘。同样，业务人员有 80% 的需求也可以直接看自助分析系统，还有 20% 可以委托数据分析人员去做深入挖掘。改进后的流程见图 4-12。

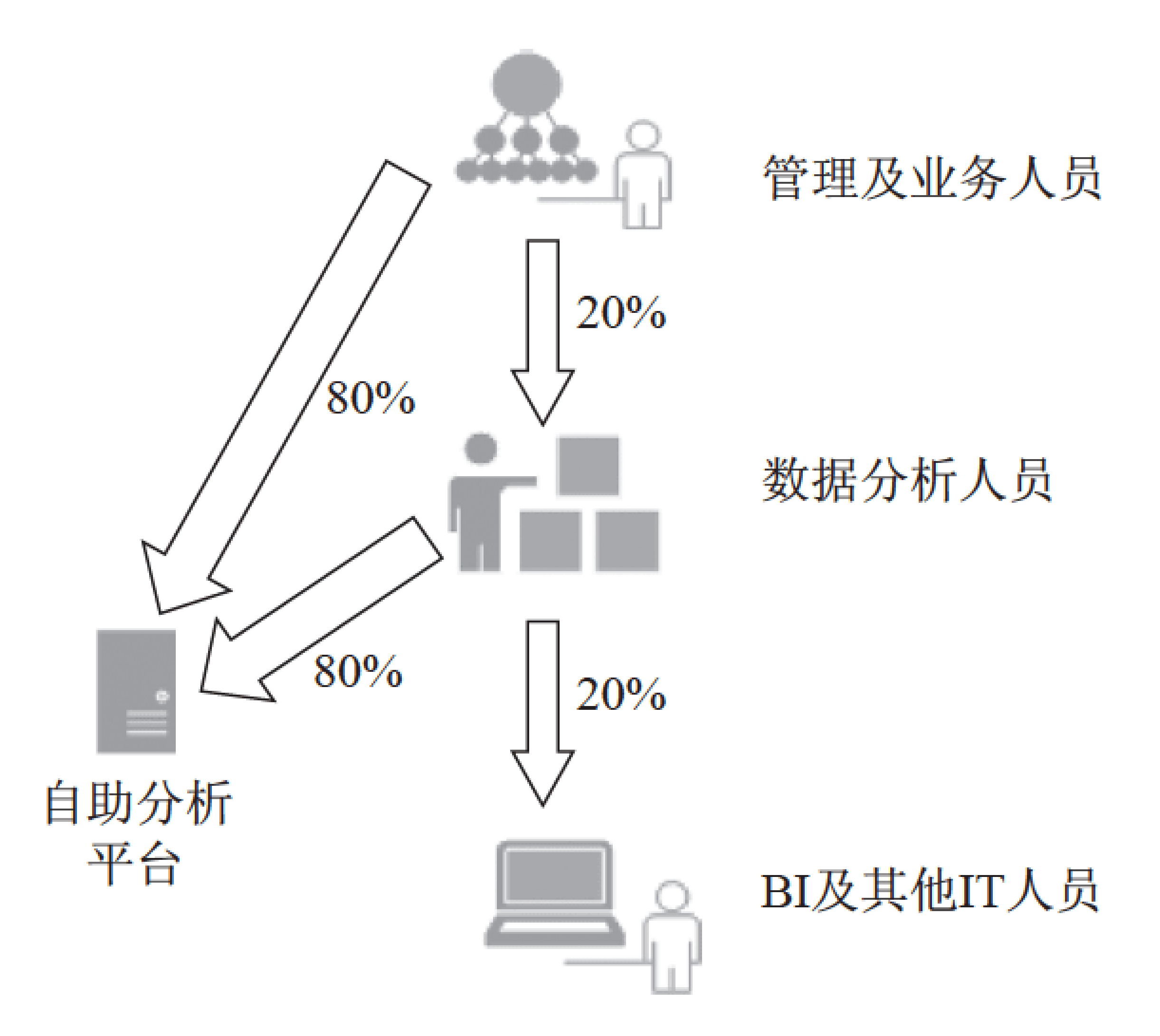


图 4-12　改进后的流程

4.4　适应数据分析的硬件

在目前，硬件的配置已经远远超过了实际的需求。现在互联网公司在发展上遇到的瓶颈主要有两个：首先由于物理的限制使它在同样面积的芯片上难以集成更多的晶体管；其次是它的高性能 CPU 功能已经超过了客户的需求，现有大多数软件无法发挥这种 CPU 的功能，也就是说，CPU 最好最贵，对软件来说在使用过程中对软件性能的影响也根本显现不出来。

但在数据分析应用中则不同，现在服务器的功能还远远不够，主要是因为大数据分析的计算量非常大，并且使用过程中出现明显的波峰波谷现象。

如果客户的请求正好命中预计算的结果，则查询的速度会非常快。但若是没有命中，它就会需要进行大量的计算。由于计算涉及很大的数据集，计算过程中会耗费很大的 CPU、内存包括硬盘，会造成在短期内占用大部分服务器的资源。因此，OLAP 的应用不适应和 OLTP 的应用共用服务器。

OLTP 的应用虽然每次访问涉及的数据比较少，每个用户对资源的需求量不大，但由于访问的人很多，众多的请求下会对资源的累计需求比较

大。不过，由于这些访问在时间上是随机的，分布比较均匀，因此对服务器资源的需求会比较均衡。

如果在一个运行比较均衡的 OLTP 的服务器上加上一个 OLAP 的应用，它会在短时间内对服务器的性能有一个虹吸的作用，受此影响，OLTP 的软件运行可能会间歇性地变慢，因此极大地影响其他信息系统的正常运行。

在硬盘的使用上，为了保证数据安全，通常情况下应采用磁盘阵列。但是，在 OLAP 应用中，用磁盘阵列主要在于考虑提高数据访问速度，因而肯定需要RAID0。至于是否需要加RAID1，看条件是否允许，应尽量加上。条件不允许时，不加未必会产生致命的影响。因为即使数据丢失了也可以重新取数，虽然会耽误一些时间，但不会造成对数据的不可逆的影响。

由于普通机械硬盘的速度与它磁头的寻址和读取速度有关，并且每个硬盘的访问速度都是有限的，如把所有的数据都放在一个硬盘上，大量的数据访问都要依赖这个硬盘的读取速度。通过 RAID0 把数据分布在不同的硬盘上，读取可以并行进行，便可大大加快数据读取的速度。

在 SSD 硬盘出现后，磁盘读写速度已经不太重要，应该尽可能采用 SSD 做硬盘。以 SAP HANA 为标志的内存数据库出现，可以把数据放在内存中，速度会更快。





第 5 章

实现数据革命

坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

国务院《促进大数据发展行动纲要》

5.1　数据革命的作用

5.1.1　对国家治理的作用

对大国的治理，一般采用联邦制和中央集权制两种统治模式。联邦制主要优点是各问其政，互相不搭界，大部分事务由各联邦主体分别来处理，只有部分的权力让渡于中央。比如美国是州的权力比较大，国家重点在国防和外交上面。而中央集权制的权力由中央主导，下面的官员全部由中央任命。

中国采用中央集权制，这跟中国的地理相对比较集中有关。秦始皇统一六国之前，各诸侯国的文字、度量衡、货币都不一致，给人们的生活和经济活动造成了很大不便，所以在秦始皇统一后建立了持续至今的中央集权制。因此，中央集权制在中国既有历史渊源，又有其合理性。但中央集权制也有很多问题：第一个是不同地方情况千差万别，如何面对个性化问题；第二个是这么庞大的机构怎样去管控。在信息时代，这种管控比信息闭塞的时代更为方便，但总的来说还是没有达到预期的目的，主要问题是没有充分利用数据。

信息化能够把所有的事物量化。现在虽然有了数据，但对数据的认识还停留在对单个数据的利用。数据的汇总还要通过统计局这类政府机构经过层层加工得到。现在统计局已经搞了直报系统，由企业直接填报、直接汇总。在数据时代，类似这种直报系统不但应该搞，而且应该大力高，但不是所有数据都可以直报的，而且统计直报会抹杀各地的个性化，挫伤地方政府的积极性。统计直报是一种理想化的信息孤岛解决模式，它不具备通用性，更多的应该建立分布式的数据仓库。

数据获取的理想方式应该是各地建立分布式系统。不同的县、市可以通过县来统一、市来统一或者省来统一，然后建立一些局部的数据仓库，通过分布式的检索，对数据进行统计汇总。数据仓库以地级市为中心建立比较合适。如果采用这种模式的话，由各个地方报数据变成了中央来取数据，就比较灵活了。

中央直取数据可以及时掌握各地的情况，不需要在出现问题的时候通过分派政府官员到各地调研的方式来解决。在政府工作的报道中，大多数是中央领导到地方考察、座谈的消息，无论是领导调研还是汇报，很少看到数据发挥的作用。口头交流或文字汇报大多是无法量化的感性东西，只有通过汇总数据才能准确掌握宏观情况。

在中国古代，为了支撑中央集权统治，建立了多种机制。一个机制是各地建立驿站，保障邮路的畅通，保证中央和地方的信息能够及时交换。另一个是官员的选拔，通过科举考试来从地方选拔优秀人才，并直接任免地方长官。

在数据时代，应该像以前重视邮路建设一样，重视数据的畅通。我们虽然关注上传下达，实际上重点是下达而不是上传，更多地重视把中央的指示通过文件的形式或者通过新闻媒体、网络媒体向下传递。

上传实际上是一种反馈。如国家作为一个物理系统，缺少一个有效的反馈机制是不利于系统完美运转的。反馈机制仅仅通过调研是远远不够的，而是需要更多通过数据。统计局统计系统的建立，初衷就是建立一个反馈系统，但比较初级：一是数据收集起来还不是很准确；二是数据加工过程中会受到一些人为的干扰，导致数据失真。而且数据经过汇总，失去了很多特征，不利于决策支持。

国家应该建立一个基于分布式数据仓库的反馈机制。

5.1.2 对国有企业改革的作用

国有企业的改革，无论方向如何、采取何种措施，都需要良好的反馈机制。这种反馈机制首先用于了解国企准确、详细的情况，其次用于评估

改革的效果。

国有企业是国家拥有全部或部分股权，国家对企业运行具有控制力的一类企业。国有企业主要集中在一些垄断行业或者基础民生行业。因为缺少市场竞争，并受体制约束，国有企业的经营和管理是一个很大的难题。在现有技术水平下，为解决这些难题，很多国家，特别是西方国家，认为竞争是唯一有效的管理方式，因此大都选择私有化。

对于各种原因不能私有化的企业，数据可以发挥非常大的作用，利用数据技术可以解决现在难以解决的许多问题。

中国的国有企业现在的主要改革方向是合并和混合所有制。所谓合并，就是减少国资委直接管理的企业，但这个企业仍然存在，如果有问题的话，问题也仍然存在，只不过把管理责任推给了另一个国企，向下推了一个层级。国企合并实际上反映了在现有技术手段下，希望通过减少企业数量，更好地掌握企业实际情况的需求。

国有企业管理的主要问题是数量多、层级多、信息不对称。国有企业的数量很大，政府无法直接监管。企业合并后，问题同样存在，只不过下放到合并的总公司的领导层，希望这些总公司的管理层比国资委更加接近企业。比如一个国企总经理原来管五个企业，现在再合并来五个企业，他管十个企业。五个企业他管得过来，十个企业他可能是管不过来。而且现在国有企业的层级非常多，有子公司、孙公司、曾孙公司，更难以管理。

大量、多层级企业的管理，需要利用数据技术来解决。利用数据钻取技术查询数据，不管多少层级都能应对。当然，数据技术无法直接进行管理，更无法直接产生效益，但可以掌握详细的生产经营信息，便于发现问题、分析问题、解决问题，以做出正确决策，后续还能够得到任何一个决策结果的及时、准确的反馈。

数据技术的利用，要求企业信息化水平比较高，采集并记录有大量数据。好在中国的国有企业信息化水平要比民营企业高。现在许多大型国有企业都统一了信息系统，比如总公司统一部署了 SAP，每个下属子公司都可以作为用户使用，不需要每个子公司独立部署一套系统。虽然这种模式对利用数据最理想，但因为 SAP 对用户应用水平和资金要求比较高，所以

不是所有国企都能普及。

下面具体介绍一下利用数据技术辅助国有大型企业决策的方案。

国有企业拥有包括财务、供应链的信息系统，作为建立数据分析用的数据仓库的数据源，这些系统的供应商和型号不限，可以是 SAP，也可以是用友、金蝶，如果是用友软件，可以是 NC，也可以是 U9、U8。

在国企总公司设立一个数据中心，把分布在不同子公司、不同格式数据库中的数据通过定时抽取的方法，放到数据仓库中。对于拥有多家国有集团公司的国资委，可以建立自己的数据仓库，集中多家集团公司的数据，也可以通过分布式系统，直接访问在各集团公司的数据。

有了数据后，分析方法不是传统的数据浏览或数据检索。这些方法只适合数据很少的小企业，大企业不行。我们必须通过统计汇总的分析方式，用汇总数据逐级钻取，才能发现问题。

假如有五个层级的企业集团，有上百个各级下属企业，怎么样发现问题呢？流程是这样的：

作为总公司的管理者在经营概要中会看到一个累计销售收入、累计销售收入的增长率，销售收入和增长率的同比增长率和环比增长率，这是最简单的数据，这个数据来源于所有的子公司、孙公司的累计。显然这种累计可以进行分解，先分解到子公司，比如说有十家子公司，看一下十家子公司的销售收入增长率怎么样。如果发现其中有一家收入下降了，可以对这家公司进行钻取，看一下这家子公司的子公司的经营数据。假如还有十家企业，这十家企业哪一家的销售收入下降了，十家企业可能九家都是好的，只有一家下降了，那么可以再对这家孙公司进行分析。我们来详细分析这家孙公司的数据，看看孙公司里面到底是哪个区域或者哪个产品种类的销售收入下降了。最后我们看到可能有一个产品销售收入下降得比较多。

通过类似以上的数据分析，我们分析众多的企业、众多的产品。如果有 100 家企业，这 100 家企业每家有 10 个产品，那就累计有 1 000 个产品。通过数据统计分析可以用非常快的速度找到出现销售收入下降的产品。找到问题所在后，再要求下面提交这个产品的分析报告，要求这个公司对产品销售收入下降的原因进行详细的分析、解释，并且提出整改措施。过一

段时间以后，我们可以再跟踪这个产品，看它的销售收入是不是有所好转。通过这样的方法就很容易对众多的企业进行管理，通过钻取从众多的数据中找到需要的数据。

5.1.3 对政府“三公”经费管理的作用

“三公”经费是指财政拨款支出安排的出国（境）费、车辆购置及运行费、公务接待费这三项经费，这是公共财政控制的核心。由于这三项费用的滥用涉及国家党政机关的公费旅游、公车消费、公款吃喝等不良行为，故为社会普遍关注。

通过数据可以寻找到解决“三公”经费的优化控制方案。

以公务车的运营费控制为例。公车运营需要油费、过路费及维修费等，由于每辆车出行的里程不同、车辆状况不同，故而其产生的费用会有很大的差异，而在这其中就会有漏洞产生，如车辆维修公司侵吞维修费等。由于漏洞巨大，所以在“三公”经费改革中政府把公车运营作为重点。目前普遍采用方法是通过发放补贴取代公车运营从而逐步取消公车。

用补贴代替公车仅仅是把不确定的费用变为确定的费用，而实际上总的费用并没有降低。由于补贴发放给个人后，用车成为个人支出，相关人员在遇到非必要性的公务活动时就可能选择不去，从而影响工作的正常开展。

取消公车、发放补贴只是解决这个问题的方法之一，在数据时代可以用数据来解决这个问题。第一，需要收集公车运营的费用。公车运营费用必须按照财务记账项目要求明确记录车牌号。第二，凭证明细中需详细列出公车的运营费用，而不是集中多个单据一起记账。第三，公开公车运营数据。“阳光是最好的防腐剂”，只要公开公车运营数据，那么所有的问题将会一目了然。

比如一个单位有 10 辆车，若把 10 辆车的运营维护费用公开，可以按照时间比较、按照维修科目比较、按照维修费比较，这其中的区别就能很明显地被看出来。

倘若放眼于一个城市，将所有的公车进行比较并由财政部门监控，就

会对每辆车的使用情况有一个宏观的了解，从而对突出事件进行控制。因为所谓的腐败只是个人的行为，所以通过比较就能促使问题暴露出来。

更为复杂的解决方案是专门开发出一个公车维护数据库件，将公车运行里程、维护费用及所有车型的公车零部件的价目表及相关内容都存入这个数据库中，在分析问题的时候就可以通过这个数据库记录的详情进行对照。

使用鹰眼技术，不需要将每个零部件进行对照，而只要财政部门将所有数据集中在一起，然后对所有车辆的费用合计进行比较。发现异常之后，可以通过数据钻取追寻其根本原因，从而对公车维修处理问题提出意见。

5.1.4 对“一带一路”战略的作用

国家的“一带一路”战略是一个非常宏伟的战略，但战略的实施注定不会一帆风顺。虽然一切刚刚开始，但已遇到一些挫折，比如在斯里兰卡投资的机场没有航班，在战乱的叙利亚收购油田，导致了很大的经济损失。造成损失的原因除了国企管理体制有一些问题以外，主要就是对沿线国家的情况缺乏充分了解。

西方国家在全球化的过程中，利用了探险家和传教士在世界各地收集的信息，虽然这些人的目的是探险或传教，但他们对当地人文、地理的了解和研究，为经济的扩展提供了很多的有价值的资料。

同样地，“一带一路”的范围是如此之大，我们需要掌握的信息也非常多。当然，如果仅仅按照原来探险家的模式去了解信息，肯定既无必要也不足够。现在这些国家都比较开放，交通也非常的方便，原来那些探险家需要花费很长时间，几年甚至十几年去完成的工作，现在坐飞机去，几天就可以完成了。

但是，仅仅拥有这些信息是不够的。我们需要更深入的信息，更深入的信息就是这些国家政治、经济方面的数据。

比如要到一个国家去投资，以斯里兰卡马塔拉 - 拉贾帕克萨国际机场为例，需要掌握斯里兰卡的经济总量、经济的分布、人口的分布、运输量等详细数据，而不仅仅是文字报告。

基于这样的思路，需要把建立“一带一路”数据库这样一个工作放到一个重要的位置上来，甚至放在所有的投资的前提上。

但是，这里面会遇到两个问题：第一个问题就是这些国家是否有这些数据；第二个问题就是它们的这些数据愿不愿意与你共享。

应对这种情况，有两个具体方案：

第一个方案是投资信息基础设施。这些国家的信息化水平比较低，这种低正好给我们带来新的投资机会，也就是说我们可以通过投资带动市场，或者说通过贷款来推动这些国家的信息化建设，也因此可以在这些国家搜集更多的数据。由于这些信息基础设施的建设是我国投资的，可以签订数据共享的协议。

第二个方案是数据交换。出于数据安全的考虑，让“一带一路”沿线国家无偿地把数据给我们共享是不可能的，只有进行数据交换，就像现在美国和它的盟国会进行情报交换一样，我们可以把中国的经济数据给这些“一带一路”沿线的国家，同时要求它们也提供相同类型数据，达到与它们进行数据交换的目的。

这两个方案应该能够得到相关国家的支持，毕竟我们出发点是善意的，主要是为了更好地投资和加强它们的建设，或者是为了把有限的资金用在最合适的地方。

“一带一路”的沿线国家的数量多，信息基础建设比较落后，我们可以在信息系统的软件、硬件、网络等各个方面提供一个完整的解决方案。这一方案如果试点成功，可以推广到其他的国家，显然既为我们“一带一路”的投资包括亚洲基础设施投资银行提供了必要的数据，同时又开拓了一个新的市场，相对于高铁和核能，可能对这些国家更为实用。

从中国视角看，是建设智慧城市，放眼全球，可以上升到智慧国家的建设。

5.1.5 对医疗改革的作用

医疗改革是许多国家面临的一个重大的问题，医药费用是社会作为公

共支出的一个非常大的陷阱、黑洞。

关于这个问题，不同国家探索出很多的方案来。比如美国医疗机构的完全私有化、英国医疗机构的全部公立。但是，这些方案在有很多优点的同时又存在一定问题，就像美国全部私有化，其医疗服务无疑是很到位的，但成本相对非常得高。英国全部公有化，可能成本控制得很好，却无法提高服务水平，造成很多的病人看病需要排队。

所以，如果医疗改革不从治疗方案的监督、医疗的运营机制和经费使用这些方面去改革，永远达不到理想的效果。就像医生寻找研究治疗癌症的方法，如果总想通过随机的发现、发明找到治疗癌症的药是很困难的，还是要通过基础的工作，从基因测序方面去系统地研究问题，才能找到攻克癌症的方法。

同理，医疗改革还需要从数据的收集、分析上去寻找解决问题的方法。只有公开医疗数据，比如一个人生病住院期间的治疗、药品费用明细，才能通过对比分析发现医疗上的问题，从而对症下药。

目前，中国的大多数医院都使用了 HIS 系统，医药数据完成数字化，而且部分做到费用公开，能够把每天的医药费用打印出清单给病人。

问题是，这种公开还是形式上的，对控制费用作用有限。首先，病人不是专家，无法对费用的高低做出判断；其次，如图 5-1 所示，提供的数据不可机读，无法用信息技术手段进行横向或纵向比较。

如果政府医疗主管部门能拥有所有医疗数据，并且可以对数据进行深入分析，则可以有针对性地改革，并得到及时反馈，医药费用的降低是比较容易做到的，改革的效果也会比较明显。

现在，国家也在建立医院数据集中式的一个平台，但在数据采集的技术手段上比较落后，基本是通过一个交换标准由医院报送数据。对数据的利用还存在瓶颈，有了数据以后如何处理，还没有成熟的手段。

政府建设的数据集中平台应该变向上报送数据为向下抽取数据。建立区域的数据仓库，可以以地级市为单位，然后把医院的数据通过网络集中到数据仓库。医院的信息系统是独立的，使用何种 HIS 系统没有限制，医院完全根据业务需要来购买及二次开发。医疗数据中心通过技术手段定期

（基本上每天一次）到 HIS 数据库中提取数据，最后集中到数据仓库中。在抽取及保存过程中，对一些数据进行转换，使医院之间可以进行横向比较。

出院病人费用明细清单

【单位：[illegible]医院】　日期范围：从 [illegible]8-31 到 [illegible]9-16

住院号：[illegible]　姓 名：[illegible]　性别：[illegible]　年龄：[illegible]　入院日期：[illegible]-31 15:09:0　科室：外四住　床位：

序号	代码	名称/规格	标准价	数量	单位	金额
14	40140	鱼精蛋白针/50mg:5ml(5支/盒)	11.2	7	支	78.4
15	40150	氨甲苯酸（止血芳酸）针/0.1g:10ml(5支/盒)	1	2	支	2
16	40151	氨甲环酸注射液/0.5g:5ml	9.775	1	支	9.78
17	44031	地塞米松针/5mg:1ml	0.559	18	支	10.07
18	44080	重组人胰岛素(普通优泌林）注射/400iu:10ml	0.15	212	iu	31.8
19	47010	维生素K1针/10mg:1ml(10支/盒)	0.437	10	支	4.35
20	47160	维生素C针/0.5g:2ml(10支/盒)	0.339	5	支	1.7
21	50071	氯化钙针/0.5g:10ml(5支/盒)	0.672	4	支	2.69
22	50086	氯化钾缓释片/0.5x12#x2板	0.23	138	片	31.74
23	50100	氯化钾针 10%/10ml	0.455	24	支	10.93
24	50110	氯化钠针/90mg:10ml	0.409	1	支	0.41
25	50111	氯化钠针(浓)/1g:10ml	0.627	40	支	25.12
26	50118	葡萄糖5%（直立式）/100ml	4.232	4	瓶	16.92
27	50121	葡萄糖 10%（直立式）/500ml	4.968	8	瓶	39.76
28	50125	葡萄糖 5%（直立式）/250ml	4.52	1	瓶	4.52
29	50160	乳酸林格（直立式）/500ml	4.474	5	瓶	22.37
30	50180	氯化钠0.9%（玻瓶）/250ml	1.756	4	瓶	7.02
31	50181	氯化钠0.9%(玻瓶）/500ml	2.328	6	瓶	13.97
32	50182	氯化钠0.9%(直立式)/250ml	4.52	1	瓶	4.52
33	50184	氯化钠0.9%（直立式）/500ml	4.796	4	瓶	19.18
34	50186	氯化钠0.9%(软袋）/50ml	3.531	9	袋	31.77
35	50187	氯化钠0.9%（直立式）/100ml	4.221	17	瓶	71.74
36	50194	注射用水*/500ml	1.99	3	瓶	5.97
37	50207	枸橼酸钾（可维加）颗粒/1.45g×20袋	1.685	42	袋	70.77
38	52010	复方氨基酸/500ml	26.416	1	瓶	26.42
39	52031	脂肪乳剂 20%/250ml	[illegible]	5	瓶	285.35
40	66241	开塞露/20ml		[illegible]	支	5.01

西药费　金额小计：5473.72

金额合计：29379.05

注：本清单仅提供参考，医保病人以出院结帐社保电脑反馈结帐单为准
自费病人以出院时的结帐总额为准

深圳市孙逸仙心血管医院 医保审核专用章

打印日期：[illegible]6 9:49:14

图 5-1　出院病人费用明细清单

国家应该建立一个分布式的数据网路，可以从各个地级市数据仓库中提取数据进行查询。数据可以提取到国家中心的数据仓库中去，也可以只查询结果，对结果进行比较，而不提取原始数据。

目前，医疗数据中心建设还有一个问题，数据标准交换设计的数据太多，医院里可能无法全部提供。实际上，这些数据不一定要马上提供，可以在使用中不断完善，制订一些政策措施，推动医院根据业务发展来完善这些数据。

比如针对某一项数据，有一个民营医院，规模比较小，业务中或信息系统中没有涉及或保存这个数据，数据中心取不到这个数据。取不到这些数据就不取，置为空值。但是，如果监管工作或者保险公司需要这个数据，没有这个数据，保险公司可能无法对这家医院的医疗项目进行报销，这时，这个医院肯定会积极展开工作，完善这个数据。如此一来，把数据的完善作为跟业务紧密关联的一种工作，而且是一个逐步完善的过程，可以降低数据收集的门槛。

至于数据存储，医院的数据都要保持最原始的，而且这个数据应该和医保、保险公司挂钩，以及时发现不合理的消费。

在地区数据仓库建立后，一些常见病的费用很容易掌握，但有些很少发生的病例成本就难以核算了。在全国联网以后，毕竟全国人员更多，在一个地区比较稀罕的病例，在全国来说，可能是一个发生次数比较频繁的病例，这样的话就有利于比较。当然，这个比较，并不是要降低医院的利润。按理来说，医院的利润应当保持不变，医生的收入也不能降低，但要核算清楚。

在数据不透明的情况下，好像医院的医疗收入比较低，但医院通过以药养医，病人的费用还是越来越高。这种不透明的方式对经济是非常有害的，实际上就是脱离了监管，医院不能合法地获取收益，最后让那些钻漏洞的人得到了好处。

5.1.6 对银行信贷风控的作用

银行传统贷款，主要关注两个方面：一个是查看企业的经营报表，另一个查看企业的抵押资产。从报表的角度看，在不诚信的环境中，报表可以作假，比如出现多套报表，一套对税务，另一套对银行；而在诚信的环境中，报表显然是静态信息，不能及时预报风险。

英国《金融时报》网站 2016 年 7 月 18 日的一篇文章《金融信息不可靠困扰中国经济》就谈了这个问题：

金融信息可靠性的全面崩溃正在加剧资本错配，后者正是中国经济效

率下滑、公司债务水平急升的根源所在。例如，想象一下，假如一家中国金融机构考虑向一家公司放贷，它可能会怎样设法找到关于这家公司的准确信息。作为首选，它可能会求助媒体报道。这样做可能被证明是欠缺考虑的。这家机构找到的文章可能已经受到了“有偿新闻”——在新闻发布会上将现金装入信封塞给记者以换取正面报道的做法——的左右。

如果媒体不可信，那中国金融机构还能指望谁呢？信用评级公司是一个显而易见的选择。

兖州煤业 (Yanzhou Coal) 曾经经营状况优良，眼下陷入困境，债务倍数飙升。该公司去年资产收益率仅勉强维持在 0.1%，中国评级机构却给出最高级别的信用评级。相比之下，国际评级机构标准普尔 (Standard & Poor’s) 给该公司的评级为“垃圾级”。类似的例子还有很多。[10]

在数据时代，只有我们把所有的企业经营数据都公开，银行才能保证贷款企业数据的准确性。首先，伪造一个数据容易，而伪造一串数据而且是相互关联的数据则非常困难了；其次，数据的公开有利于银行动态监控企业的状况，能比较准确地掌握企业的运营情况和资金流向，最大限度地规避风险，比单看报表要可靠得多。

从资产抵押方面看，资产抵押相对而言是一种比较笨的方法，在出现坏账时，资产变现十分困难。抵押资产增加了企业贷款的门槛，有些做得好的企业并不一定拥有很多可抵押资产，这样容易把优良的企业拒之门外。因此，基于数据和公司供应链及现金流的监控，能够部分解决中小企业贷款难问题。

从供应链贷款的角度来看，通过对核心企业采购数据的把控，可以对属于核心企业的供应商提供供应链贷款。原来供应链贷款可能需要对核心企业提出很多要求，甚至部署专门的信息系统，不然难以获取这类数据，而采用现在的数据技术，如果能直接从它的 ERP 中获取数据，就可以极大地减轻核心企业的负担。由于核心企业不需要额外做什么事情，只需要开放 ERP 的数据访问权限即可，从而大大降低了供应链融资的推广门槛。

下面详细介绍一下利用数据进行供应链融资的方法。

银行的贷款常常需要抵押，信用贷款比较少，基于信用和抵押之间现在比较重视供应链的融资，但供应链的融资涉及核心企业数据的获取。由于企业供应链的融资受惠的主要是核心企业的供应商而不是自身，核心企业对此兴趣不是很大。有的银行觉得，可以通过给核心企业分成来提高核心企业的积极性，但实际上每个核心企业都有自己的盈利模式，如果它过于关注供应链融资的分成，就变成多种经营了，因此公司层面不会很重视。如果供应链融资还要求核心企业为此专门配置一套信息系统，那么门槛就更高了，因为一家企业使用一套信息系统的成本是非常高的，不但要有人维护还要有人输入数据。虽然这其中会对它产生价值，但毕竟企业最关注的是自己的核心业务。因此，现在最好的方法是对核心企业不提什么要求，而是利用它现有的信息系统的数据来实现供应链融资。

众所周知，能成为一个核心企业，肯定是经济效益和管理水平比较高的企业，这种企业一般都有 ERP 系统，而 ERP 系统里就包括它的采购方的采购信息、发货信息及它的付款信息，所以只要能从 ERP 中抽取数据，就可以监控到它的供应商的货物和资金的来往情况，达到保障资金安全的目的。

供应方的融资有两种方式。一种方式是希望由核心企业来担保，也就是说核心企业付给供应商的钱受到银行的监控，这样即使供应商企业有问题，这笔钱可以被银行扣留，但这种方式对核心企业提出了新的要求，增加了业务开展的门槛。另外一种方式则比较简单，不需要管资金，只要监控数据即可控制风险。

实际上，任何一家企业出现经营上的困境不是一两天的事情，它有一个酝酿的过程，只要你监控核心企业的数据，这个过程发展趋势是非常明显的。比如说一个供应商如果经营状况在恶化的话，核心企业会逐渐减少对它的采购，而供应商的发货周期也会延长，核心企业在对供应商的付款上也会出现异常，所有这些信息实际上核心企业的采购人员心里是知道的。

既然我们要采用大数据去实现供应链融资，就不需要经常要求核心企业采购人员评估这些供应商，而可以通过数据的监控发现供应商的数据异常。比如说，核心企业对该供应商的同比采购数量下降或者是环比下降很

多，就应该立即引起重视，然后在贷款的清收上或下一笔贷款的发放上采取谨慎的措施；或者在发现数据异常以后，马上向核心企业的采购人员了解信息，得到一些更准确的反馈，防止误判。

银行可以请求核心企业对 ERP 数据进行开放，银行通过远程看到数据分析的结果。为了安全起见，可以只开放采购和应付款的数据，对企业的其他数据一概屏蔽。核心企业可以通过对数据权限的设置来防止其他数据的泄密。另外，也可以由银行提供名单，核心企业对名单上的供应商开放数据，以此保证数据的安全。一般来说，企业对自己销售的客户和销售的数据比较关注，对安全性比较重视，但对采购的数据相对而言不太重视，除非一些核心的供应商，但现在这种供应商在企业中的数量还是比较少的，大部分都是一些比较大众的供应商。

5.1.7　对降低社会成本的作用

中国要跨越中等收入陷阱，进入发达国家行列，必须在提高人力成本的前提下降低其他成本。中国过去三十多年的发展，是以廉价的人力成本掩盖了其他众多高昂的成本。以美国为首的西方发达国家，实际上在人力资本外，也拥有很多的成本优势，中国不做深层次的改革这将成为跨入发达国家的主要障碍。这种改革的成功除改革的意愿、改革思路、改革执行力外，更需要对改革进展和结果的掌握。

是否要改革，如何改革，表面上体现的是政治意愿，实际上通过数据可以准确反映改革的必要性和改革的效果。如果改革仅靠专家和媒体的呼吁，由于他们得到的信息不完整，而代表被改革者利益的业内人士掌握的数据比较具体，所以开会讨论时，只要他们摆几个具体问题而又没有对策，决策层就会犹豫，导致改革停摆。

发起改革的理由和改革后的效果，都要有基于数据有说服力的证据，才能得出社会共识。比如说垄断，需要公开垄断企业的运营数据：它的成本发生在哪里？利润在哪里？效益在哪里？人均效率在哪里？对数据进行分析，把数据和国内外同行进行比较，从而发现问题在哪里？哪里成本偏高？

比如一个城市的自来水公司，它的供水成本有人反映比较高，等到具体研究降价的时候，自来水公司在会议上会拿出证据证明它是微利甚至是亏损，这样将无法说服领导对价格进行调整。如果开听证会，市民代表手里没有数据，只能听从一面之词，最后听证会都开成涨价会。如果这个自来水公司能够把它的运营成本公开，实际运营的盈利情况公开，关键成本开支是否合理，将会一目了然。

如果全国所有城市的运营成本都公开，不同的地区显然会存在差异，比如采购价格的差异。但通过横向比较，一方面能明显地看出哪些城市成本支出不合理，揭开那些虚高运营成本的面纱；另一方面也能够让运营公司看到自身的不足从而进行调整，使全国的供水成本均衡化，不给个别企业获取暴利的可能性。国家也因此能制定一个公用事业平均的毛利标准，让企业有一个合理利润，保证其正常运营。

数据公开还可以揭露垄断行业的费用在整个社会成本中的比例。通过成本的公开，可以发现商场一件衣服的成本构成、可以分析出地价在这其中累积的占比，看到政府由于累积的收费最后进入财政的钱的占比。

通过这种对复杂产业链的跟踪揭露它的主要成本，相关问题凸显，并且也很容易与社会形成共识。在这种的情况下再去谈改革，比较容易得到大家的认可。

由于这些数据与改革的措施有很大关联，可以明确地测算出改革可以取到的效果，最后可以根据数据看出改革的结果和改革的预期是否相符。

比如，现在绝大多数人对高房价持反对态度，但有些人认为高房价带动了产业链的发展，促进了就业。这里有一个主要的研究课题是：高房价带来的就业和贡献与它带来的危害之间的比较。这是一个很复杂的课题，但通过数据的研究可能会发现实际上高房价的危害远远大于它带来的对社会的效益，因为我们每个人在生活消费中都在为高房价付款。

如果研究美国经济，可以发现这样一个链条：物价低是因为商业房租低，房租低是因为汽车使用成本低（没有或低过路费、低汽油价格）。原来我们总认为美国是“汽车上的社会”，它的低油价会造成资源的浪费，却没有看到它对降低成本的作用。

5.1.8 对防止欺诈上市的作用

建立健康发展的证券市场，推动企业直接融资，对中国实体经济的发展至关重要。信息是证券市场赖以生存的基础。为消除信息不对称对投资者造成的侵害，各国均制定了相关的法律法规，强制信息持有者进行公开，并辅之以相应的法律责任。欺诈上市的行为在证券市场上屡禁不止，严重损害了市场的正常运行和投资者的利益。

在监管层的券商风控新规要求下，一旦出现欺诈发行或信披违规，保荐券商或将承担较大的连带责任，因此各券商都很重视风险管理。

风险管理的基本程序是风险识别、风险评估、风险应对，而三者都是建立在信息搜集的基础之上的。如果无法搜集到有效的信息，整个风险管理体系就会失效。

尽职调查是 IPO 项目风险管理的关键，尽职调查是整个 IPO 业务风险集聚的环节，风险点众多。财务尽职调查是尽职调查的关键。

尽职调查的重大风险包括持续经营风险、税收风险、关联方交易风险、舞弊风险、战略经营风险、独立性风险等。

财务尽职调查要识别出虚增收入、虚增利润、关联交易等风险点。

现在的尽职调查因为无法通过对企业的财务凭证数据、ERP 进销存数据进行详细分析，很多工作只能手工进行，从外围入手，不但数据准确性差，而且耗时太长，一般都要半年左右。

在数据时代，可以实现从宏观到微观，即从行业到业务，再到财务的分析。行业分析数据来自对已上市的同行公司的财务数据分析，业务分析数据来自企业 ERP 数据，可对企业供应链的经营数据分析，财务分析数据来自财务软件。这样，可以减轻尽职调查工作的难度和强度，减少尽职调查的人员数量，大幅缩短尽职调查时间，从而收集企业更多的有效信息，识别更多可能的风险点。

在拥有这些数据后，具体如何方便地找到风险点呢？

对于利用虚假交易虚增收入，可以：①通过对合同金额的排名，和同比排名，找到异常交易；②通过分月、分日销售金额对比，在应该比较均

匀的金额中，发现异常交易的存在，再通过数据钻取和销售明细找到异常交易。虚增收入可能体现在某几日收入太高；③通过对平均最高单价分析，发现抬高单价的异常交易；④通过对销售业务流程各环节对应关系，如出库数量、发票金额、应收款金额、收款金额的关联性分析，发现异常交易；⑤对企业提交的利润表中营业收入和自动从财务凭证中生成利润表营业收入进行比较。

对于虚构利润，可以：①将财务分析中主营业务毛利和经营分析中的毛利进行比较；②比较同行业上市公司的毛利率、存货周转率、财务费用率等数据，发现隐藏财务费用等的线索；③按时间维度观察对主要客户或供应商的销售和采购平均单价，是否和其他供应商的平均单价比较过高或过低。

5.2 数据革命的后果

大数据成为提升政府治理能力的新途径。大数据应用能够揭示传统技术方式难以展现的关联关系，推动政府数据开放共享，促进社会事业数据融合和资源整合，将极大提升政府整体数据分析能力，为有效处理复杂社会问题提供新的手段。建立“用数据说话、用数据决策、用数据管理、用数据创新”的管理机制，实现基于数据的科学决策，将推动政府管理理念和社会治理模式进步，加快建设与社会主义市场经济体制和中国特色社会主义事业发展相适应的法治政府、创新政府、廉洁政府和服务型政府，逐步实现政府治理能力现代化。

国务院《促进大数据发展行动纲要》

5.2.1 竞争机制的替代

西方经济几百年来得到迅速发展，这与它的市场经济制度设计有关。

这种设计包括有限公司的设计、市场竞争机制的设计，都是非常有效的手段。应该说，市场竞争是在当时的技术条件下最佳的运作模式。

到了数据时代，这种模式会不会有变化呢？应该说会发生变化，市场竞争将不再是资源配置的最佳方式，起码不是唯一的方式。

市场竞争有很多的局限性，而且为了避免这种局限性，人们也做了很多努力，比如说通过兼并来避免恶性竞争。两个企业如果处于竞争状态，虽然价格很低，但因为市场部是重复设置的，对社会来说是浪费的。企业通过兼并来解决市场竞争低下的问题，也就是通过兼并把两个市场部合成一个市场部，保留不同的产品部，以提高效率。

但兼并又可能形成垄断，为避免垄断带来的高价格，国家又制定了反垄断法。

市场竞争是有成本的。现在有很多的广告，一类广告是以产品信息为目的，这是必须的；另一类仅为竞争的目的，就是浪费。

能不能有一种既能达到竞争的目的，又能避免竞争的浪费的方法？利用数据可以做到。通过数据的公开和数据的分享，可以减少竞争的负面作用。

下面分析一下竞争有效性的原因。由于信息不透明，消费者很难知道一个产品的真正的、合理的成本是多少，所以无法对售价提出要求。只有通过竞争，也就是说另外一个厂家生产同样的产品，它的价格如果能降低的话，就能证明原来的厂家价格也能降低，消费者因此知道了一个产品可以达到的最低价格。由于很多消费者转为选择新厂家、新产品，迫使老厂家降低售价。如果老厂家确实由于生产率低下无法降价，就只有破产。

如果有一个非常公开的成本架构，消费者可以知道它的真实成本，也可以有效地降低产品价格，产生跟竞争类似的效果。厂家知道业内最优的成本构成，也可以及时发现问题，提高效率，或及时转移产能。现在一些事业单位或垄断企业由于它的成本没有公开，也没有合理竞争，会受到公众的压力，如果它的成本结构能够公开，而且利润合理的话，压力会减小，如果成本确实过高，则通过公众的压力或者政治上的压力也可以迫使它降低成本，从而降低社会成本。

5.2.2　计划经济和市场经济的融合

计划经济由政府对经济进行集中掌控，政府占有、经营土地和经济生产的资本、资料，在 20 世纪 30 年代—20 世纪 80 年代，都认为计划经济是比市场经济更有效率、更公平的经济生产形态，由政府集中控制经济管理和企业决策比分散的经济秩序更有效率。

市场经济由私人占有生产资料，私人组织生产活动，生产经营活动是建立在提供和利用零碎而分散的信息基础上，价格和利润传递着各种商品和服务的相对供需状况。

现在一个国家的经济制度基本介于计划经济和市场经济之间。将倾向计划经济的政党称为左翼政党，倾向市场经济的政党称为右翼政党，还有很多持中间立场的政党。

一个国家的经济最好要像汽车一样，行驶在路中央，而实际上不是偏左，就是偏右。在西方国家，一个左翼政党执政久了，就会偏左，如果经济出了问题，选民就会拥护右翼政党上台，右翼政党肯定偏右，开始正好纠正了“左倾”的错误，经济会不断变好，但慢慢就偏离了中间，开始问题不大，但慢慢就出了问题，要想纠正右倾的错误，只有让左翼政党上台。如此往复，经济就像汽车，不断向前行驶。

在第二次世界大战结束后，由于战时经济被管制，生产采用计划经济，加上苏联由于实行计划经济表现出整体实力，所以很多欧洲国家都倾向于计划经济，“左倾”政府得势。到 20 世纪 80 年代，由于西欧经济衰退以及苏联机制暴露出的问题，致使英国撒切尔夫人和美国里根总统进行了私有化和放松管制的改革，转向右倾。

计划经济的优点是由政府掌握生产资料，通过计划指导生产，优势是可以科学规划，总体调度，避免私人无序竞争，劣势是由于经济的复杂性，决策者知识的局限性，导致生产效率的实际下降。哈耶克认为知识分散在所有人的头脑中，这些零散的知识不可能被汇集到一个人的头脑中。[13]

市场经济由资本家根据价格和利润信号组织生产，但由于信息的不对称和消费者的不理性，这种信号传导常常会出错，因此市场经济也存在大

量资源错配，经济危机就是调整这种资源错配的机制，但经济危机的发生也导致大量社会资源的浪费，甚至导致政治危机。

寻求计划经济和市场经济之间的中间道路，是很多知识分子和政党的追求，但在数据时代之前，由于缺乏有效技术手段，并没有找到合适的解决方法，而数据革命，为计划经济和市场经济的融合创造了条件。

数据革命将收集海量数据，并将这些数据提供给市场参与者，而决策支持系统的建设使这些参与者可以利用各自的专业知识理解数据中包含的信息，从而做出正确的决策。有了这些信息，既可避免信息的不对称，又可避免个人的不理性，彻底解决了市场经济的短板，从而把计划经济的长处利用起来。

这样，计划经济不再是中央计划制订部门部分人拟订的计划，而是一种分布式的有很多人参与的计划，比市场经济中价格及利润信号更为准确、及时、全面。

5.2.3 经济危机的消除

经济危机不是坏事，当经济存在泡沫的时候，危机是有利于返回经济均衡状态的一个工具。如果像日本逝去的 20 年一样，在泡沫积累以后没有及时去消除，亏损企业没有退出市场，而是靠银行输血维持，造成的损失更大。经济危机应该是数据时代以前的一个产物，在数据革命成熟以后，这种危机应该不会存在。

危机产生的一个原因就是无法获取权威的、准确的数据，无法判断在某一个领域投资是否恰当。

一个领域的投资总是以下四种状态之一：一是不足、二是恰好、三是过剩、四是泡沫。投资处在恰好状态的时间是很短的，大多时间不是不足，就是过剩，过度过剩状态就是泡沫。然而，目前是什么状态，是否已经投资过剩成为泡沫，人们通常无法判断。由于大家得知的信息比较片面，容易造成大家都向一个领域投资，从而产生泡沫。

在数据时代，由于信息多样化，大家获取信息的成本比较低，而且只

要有一定的专业知识，花一定的功夫就可以准确地了解这方面的信息，所以大家的投资就会比较分散，而且效益比较高，不会集中涌在泡沫所在的区域，这样就能够消除经济危机产生的根源，即投资的泡沫。

5.3 数据革命后的技术

5.3.1 以数据检索为主的搜索引擎

以谷歌、百度为代表的搜索引擎，无论是技术还是商业模式都比较成熟，下一步该如何发展呢？

现在搜索引擎主要是搜索网上的网页，就是对非结构化的数据进行搜索，然后跳转到相关的网页上。

从数据含金量的指标分析，网页数据的含金量比较低。在数据时代到来以后，许多互联网资源应该转变为数据资源，这样含金量会大大提升。当大量的资源都是数据的时候，现在的搜索引擎已经不能满足要求了，这就需要对搜索引擎进行升级。

现在的百度可以对 Excel、PDF 等格式文件进行搜索，产品名称叫百度文库。这些对 DOC 文件、TXT 文件、PDF 文件、PPT 文件或 XLS 文件的搜索，比一般的网页搜索引擎更接近于数据搜索。

数据搜索应该是以 SQL 语言为标准语言的一种搜索，可以直接输入 SQL 语言进行查询，也可以输入一些关键词组合成 SQL 语言进行查询。查询结果输出数据，此外，还要包含大量对数据的解释，而不仅仅是数据本身。

虽然雅虎开创的分类检索已经被搜索引擎替代，但数据引擎对数据库的分类还可以用到，它可以作为搜索引擎的一种补充。因为对数据的使用方式主要是读取，需要利用类似照相机镜头的变焦功能完成从宏观到微观，或微观到宏观的一种自由切换，跟地图检索的方式非常接近。这种情况下，

通过分类，分成大类、小类来逐级寻找数据，应该比较符合大家对数据读取的习惯。

数据搜索引擎比现在的网页搜索引擎重要性要高。搜索引擎升级以后，应该和现有搜索引擎兼容。

5.3.2 基于数据的云服务

云计算已经得到非常大的发展，很多公司都推出了云计算的平台，云计算的概念也逐步为大家所接受。

云计算实际上是模仿了第二次工业革命的方法，电力的供应是集中在发电厂，通过输电线路输到各个企业和家庭，用户只要花很少的钱，就可以共享这么庞大的发电机的投资成果。

云计算通过集中大量的服务器，将服务器拥有的计算能力通过网络共享，让用户根据需要来付费。从数据革命的未来来衡量，云计算目前做到这一步还是初级阶段。和成熟的电力供应相比，云计算服务有一个明显的缺陷。我们知道，电力线路输送的电力产品是一个标准化的产品，用户只要通过一个标准插头接入电力线网络就可以获取电力。而在云计算环境中，通过互联网接入云计算平台后，所获得的数据千差万别，无法直接使用。

利用现有的云计算功能，每个人只能读自己保存在云计算平台上的数据，难以看到或看懂别人的数据。虽然云平台也保存一下公用格式文件，可以通过共享看到，但显然跟电力供应的原理有很大的不同。发电企业所发的电与某一个客户的规格和需求无关，是根据自己的计划来发电，而且同样的电力可以在不同的用户之间进行任意调配。

现在带数据的云只能分割成一块一块的，为每一个用户定制，类似一个电厂有好多发电机，每个用户只能租其中一个发电机，在他不用的时候发电机不可以给别人用。

数据时代的云计算平台主要存储一些公用的数据，数据存储格式应该是标准的、固定的，其他人可以方便认识和共享。

现在云计算供应商一般分为三个类型。第一个是 IaaS，提供以硬件为基础的基础设施；第二个是 PaaS，提供一个公用的软件开发平台；第三个是 SaaS，提供基于软件的服务。

数据时代会出现 DaaS 供应商，提供数据服务。随着 DaaS 服务内容的标准化、服务对象越来越多，这种服务应该成为云计算的主流。慢慢地，其他云计算的服务都成为它的附属设施或者服务保障。

为什么 DaaS 会在这几种服务中脱颖而出呢？因为 IaaS 的缺点是不面向最终客户，因此它的市场容易为下游的供应商所控制；SaaS 由于软件的功能比较局限，用户面比较窄，所以规模不会很大；PaaS 本身的规模不大，处于中间层次，既可以被 SaaS 的厂商替代也可以被 IaaS 的厂商所替代，所以它更不会形成一种独立的竞争实力。但是，DaaS 因为提供的产品比较标准化，服务面比较广，所以它会成为以后云计算服务商的主体。也许刚开始，DaaS 可能会利用 IaaS 的基础设施，并且从 SaaS 厂商那里获取数据作为起步，但随着规模的扩大，其他的厂商都无法独立生存，最后被 DaaS 的厂商合并吸收。

5.3.3 可以检索数据的浏览器

现在的浏览器已成为很多人进入互联网的入口，曾有过代替操作系统的趋势。

但是，相对于全球的数据量来说，通过浏览器看到的只是很小的一部分信息，只有网络服务器上以 HTML 语言作为标记的标准文本才能被浏览器访问到。一般的网页都比较简单，不含有大量的数据和信息。也就是说，虽然我们可以通过浏览器浏览全世界范围的网页，但这些网页所含的信息量在世界上是很少的。

同样地，如果我们用 Google 及类似的搜索引擎仅仅搜索网页，实际上也仅搜索了很小部分的信息。更多的信息应该是数据。为了支持数据的搜索浏览器就需要改进，目前的 HTML 版本是 5.0，也许在后面的 HTML 版本中会加入对数据的支持。数据最好采用 SQL 语言访问，但可以加一些标

记由浏览器解读。

虽然很多数据可以通过编程，以网页的方式展示，但毕竟这些数据和网页还是捆绑在一起的。网页是开发一个程序，由前台和后台构成。未来的发展目标是数据和程序分离，也就是可以不通过网页程序，而是通过浏览器就可以直接访问数据。



第 6 章
工业数据革命

引文案例

张总是一家大型上市公司总经理。

2030年秋季的一个普通日子，晴空万里，张总怀着愉悦的心情，脚步轻快地来到办公室。

张总的办公室一面是落地玻璃，可以看到这个城市CBD的全景，一面是一个矩形液晶屏幕，屏幕前以U字形围着一圈高级沙发，沙发前有一个可以移动的小车，小车上有一个触摸屏，大概有40英寸，触摸屏显示的内容可以同时显示在液晶屏幕上。张总的办公桌在沙发后面，桌面上也有一个32英寸左右的4K显示器，但不像普通显示器横放，而是竖放着。

张总坐到办公桌前，打开电脑，进入公司内部网站，输入自己的用户名和密码，首先呈现在眼前的是一个综合显示经营数据的仪表板，上面展示出昨天晚上更新的最新数据，显示到昨天下班为止公司主要的经营数据，仪表板上用多个像汽车速度仪一样的小仪表盘列出累计销售数量、累计销售金额、累计毛利、毛利率、应收款余额、累计采购订单金额、累计采购入库数量、应付款余额、库存数量，除余额和毛利率外，这些数据都是年初至今的合计，也就是从今年1月1日开始到昨天的累计数据，并且用指针显示和去年同期的比较。

张总重点关注一下每个指标的指针，这个指针可以快速看到和去年同期的比较情况，如果指针偏左，表明数据低于去年，如果偏右，则表明数据高于去年。为便于对经营状况的快速识别，仪表盘用颜色醒目地分为两个区，绿区和红区。如果指针在绿区，表明数据指标正常，如果在红区，则指标异常，需要引起关注。根据指标不同，有的仪表盘绿区在左红区在右，有的则绿区在右红区在左。比如，累计销售数量、累计销售金额、累计毛利等指标是越大越好，所以绿区在右，应收款余额、应付款余额和库存余额等指标则是越小越好，所以绿区在左边。

张总用眼睛扫了一下所有仪表盘，发现大部分都在绿区，只有应收款余额在中间偏右一点。“最近应收款清收放松一下，马上余额就增加了，要找销售部算账”，张总自言自语地说。

张总没有马上叫销售部经理过来，而是再仔细地对数据进行了分析。

他在仪表板上看了一下应收款余额同比增长率最高的10个客户名单和每家的应收款金额，发现排名第一的增长率特别高，是第二名的两倍。

他进入应收款分析主题，按业务维度查询应收款余额，发现销售2部合计余额最大，再看同比，和去年同期相比，也是销售2部同比增长率最高。到底是销售2部所有业务员的应收款余额增加，还是某个人增加呢？他用鼠标点击销售2部，进入销售2部所有业务员的数据比较页面，发现一个名叫李明的业务员应收款余额同比增长率最高。再点击李明，把业务员维度的值锁定为李明，接着进入历史维度和客户维度，发现两个重要线索：

（1）李明的应收款余额是这个月刚高起来的，原来是正常的；

（2）应收款余额增长最高的客户就是李明的客户。

于是，张总在掌握了这些详细信息后，给销售部经理打个电话，让他和销售2部经理、李明一起到他办公室。

几个人到达办公室后，坐在液晶屏前面的沙发上，销售部经理熟练地打开液晶屏幕和移动触摸屏，屏幕上显示出和张总桌上电脑相同的页面。张总从办公桌前走过来，坐在中间的沙发上，操作触摸屏把刚才自己在桌上电脑分析的流程走一遍，要求销售部的这几个人解释一下应收款余额增加的原因，并要求整个销售部根据这个情况作出对策，防止其他客户的应收款余额也增加。

等这几个人离开办公室，张总回到办公椅子上，把椅子转向落地窗，心里庆幸有这个先进的决策支持系统。如果是传统管理模式，首先自己不会发现这种情况，只有在财务部门汇报的时候才会发现，然后马上就会把销售部门的人找过来，要求销售部门寻找原因进行整改，然后到下个月的财务报告出来的时候，才能发现销售部门的整改是否到位，是继续要求整改还是把这个问题过掉。也就是说，如果没有一个详细的数据分析，第一个他不可能自己发现这个问题，而需要财务部门发现问题后来汇报；第二

个是发现问题之后自己解决不了，找不到问题所在，所以只有责令下属去找出问题，下属是否能找出问题他也无从知道；第三个是反馈的时间非常长，必须等下个月的数据出来以后才能发现。有的时候找到问题的根源后，是否能找到解决问题的方法，最后能否得到认真执行，完全依赖于销售部门负责人的能力和信任，而他自己常常是无能为力。这个在一些管理理论上美其名曰“结果导向”，也就是说只要应收款余额正常了就好，至于下属是怎么采取措施的就不管了。传统的管理类似“黑箱管理”，是在缺乏数据的情况下的无奈之举，但被一些理论家提到一定的高度以后，很多人把它当作一个法宝。

张总现在对基于决策支持系统的管理新方法已经驾轻就熟，在有了详细的数据以后，很多工作不必依赖下属，而是通过数据分析就可以发现出现这种情况的原因，比如说可以马上进入应收款的数据分析这个主题里面去，看一下今年应收款余额的按时间整体走势，还可以看一下同比的走势，然后从这里面发现是哪些区域、哪些客户或者是哪一个销售人员数据上升。一般情况下，不可能这几个情况同时都发生，更多的可能是某个区域上升了，一般的一个区域是由几个销售人员负责的，我们就会发现可能是这几个甚至是某一个客户应收款的账期超过较多，甚至只是一个大客户。在锁定这个大客户以后可以看一下这个大客户的应收款的历史情况，有可能这个大客户最近的应收款的时间明显拉长，当然也有可能整个区域的客户都产生问题。这种情况下就可以有针对性地叫销售部门的主管和分管这一区域的组长甚至是营销人员过来开会，要求他们对这种现象进行分析，找出原因，也许可能需要由销售人员到客户现场去拜访找出原因，也许这个工厂由于经营不善可能要接近破产从而导致了应收款的账期拉长，这种情况下要把该客户列入警示名单，谨慎地发货，缩短账期，采取一定的措施加紧对现有应收款的催收，减少新的应收款避免出现坏账。所以，通过数据分析可以发现工作进行中隐藏的问题，直接分析出问题发生的原因，找到解决问题的方法。

上午其他时间无事，张总在脑子里做了一些比较前瞻性的思考，又到研究所关注一下新产品的研发进度，到几个部门找几个高薪挖过来的人聊

聊天，看到他们的状态都不错。

转眼到下午，是月度例行会议的时间，所有的公司高层都集中到会议室里进行例行的经营分析会，对公司最近一个月的经营情况进行分析，看看出现了哪些问题。

会议室除通常的会议桌外，一面墙上也是类似张总办公室的大型液晶屏幕，但屏幕前已没有移动的触摸屏，而是在每个座位前都放了一个平板电脑，大小只有 21 英寸。

会议按惯例由各个部门汇报情况，流程是先口头汇报，再用 PPT 显示相关数据的统计图形，这些图形都是从决策支持系统上截图下来的，如果有问题就打开决策支持系统从不同维度对数据进行研究，找出问题的原因。

在财务部汇报时，不出所料地提出应收款余额增加的问题。财务部经理按照时间维度在一张图上展示销售金额、收款金额、应收款余额的走势折线图；另一张图是同样指标与去年同期相比的同比折线图。在图上发现，从上个月开始，应收款余额增加了，原因是虽然销售额增加，但收款金额下降，因此应收款余额上升的幅度高于销售增加的幅度，这从同比的折线图上看起来更一目了然。

根据财务部发现的问题，结合上午张总处理的特例，销售部经理分析了出现这种情况的原因，他们准备采取的措施和预计改善的时间。

这样有效率的公司例会，在以前是不可想象的。张总清晰地记得，以前财务部门在发现有问题以后，肯定是由销售部门笼统解释一下这样的原因，并且提出一些整改的措施。如果销售部经理原来已经发现这个问题或者财务部门私下已经跟他交流过，点出这个问题，他可能就会找出问题所在，比如说，由于某个大客户应收款清收的问题导致这样的情况。他会把这个情况汇报一下，并且通过几个图形把展示出他们是如何找出这个问题，来证明确实是导致这种情况的原因。如果销售部门原来没有发现这个问题，刚刚得知的话，销售部经理在会议上无法提出具体解决方案，只能回去研究这个问题，找出原因并且在下次会议上提出方法或者直接整改以后把结果汇报一下。

例会结束后，张总回到办公室，泡一杯咖啡，坐下来打开邮箱，看看有什么邮件需要处理。他看到南方子公司 A 公司审计部的李经理的一封邮

件，汇报一下最近的审计成果，主要是告诉张总，他通过比较分析，发现A公司的库存余额同比增长率比较高，并把自己找到的问题原因和解决问题的建议也写在邮件中，邮件中还附了一张柱形图，显示出A公司库存余额同比增长率的时间走势情况，并告诉张总他已给张总发了书签，张总自己可以找到这幅图并做深入分析。

张总在决策支持系统中找到李经理发给自己的书签，在库存数据分析主题中打开与邮件中相同的统计图，发现确实存在这个被自己忽略的问题。他通过对时间维度、仓库维度、存货维度的单独与综合分析，验证一下李经理的分析结果，基本同意李经理的判断和建议，但在和生产部门的协调上提出一个补充意见。随后，综合李经理意见和自己意见，写了一个邮件发给了A公司总经理，并抄送李经理。

不知不觉，到了下班时间，张总关掉电脑，拿起公文包，准时下班。准时下班，对张总是常态而不是一种奢侈，因为他总是能够及时发现生产经营中的问题迹象，在问题变得严重前预先处理了，很少发生突如其来的状况。虽然张总通过智能手机也可以使用决策支持系统，但他很少在下班后使用，大部分是在出差时使用。

6.1 智能制造首先要解决数据问题

信息时代结束以后，有一种说法，认为下一个时代是智能制造的时代。

智能制造是一种制造业发展的目标，现在比较流行的一些概念和技术，比如工业互联网、工业4.0、机器人，都是为智能制造服务的。

那么，智能制造时代有哪些特点呢？既然是智能，首先是智能系统，可以说是一个人体系统的简化版，因为人体是智能系统的最高境界。

机器人发展的最高目标是模仿人，现在工业上很多机器人是一些机器臂或者说只是模仿人局部的一个功能，并不能模仿一个整体的人。

人的神经系统由中枢神经系统和周边神经系统组成，中枢神经系统由

脑和脊椎组成，神经系统由数以亿计的细胞（神经元）组成，脑发出的指令通过神经元迅速传递到身体，将身体接收的信息传递给大脑。

一个人完整的智能控制流程是怎样的呢？简单来说，人体的智能活动由三部分组成：首先是通过各种信息的输入渠道感知外部的信息，包括视觉感知图像信息、耳朵感觉声音信息和其他的味觉、嗅觉等器官来感知其他信息，然后把这些信息传到大脑，大脑再对这些信息进行决策，决策的结果再传递到人体的四肢形成动作。这就是人体智能的基本流程。

显然，这个流程是一个循环，而且是一种闭环结构，一个不断地循环往复的过程，人体系统根据动作产生新的信息，并得到反馈，来不断对动作进行调整，其中的核心是神经系统。神经系统有上传信号和下传信号两种，上传信号就是把收集到的信息上传到大脑，下传信号就是把大脑的决策传递到四肢进行动作。

神经系统在人体的智能系统起到非常重要的作用，进一步分解来说，人体的活动需要有信息的获取、决策的制定和动作的进行这三个部分。

仔细分析一下能够发现，如果不考虑人工智能的决策，也不考虑执行，仅仅只考虑信息的采集，能不能把各个制造系统各部件的信号采集输送到控制中心，让中心的中央决策系统能够看懂这些信号的含义，这个问题现在并没有解决，这也是数据革命的核心内容。

如果说大家认可数据革命，认为数据革命是一项必须的工作，而且认为数据革命是非常复杂的，需要人们花很多时间去实现这样一项工作的话，那么大家会想到智能制造还很遥远。

目前，首先要进行的还是数据革命，即先解决数据的采集、传输和识别，至于根据这个识别如何做出相应的动作，现在人工智能研究发展很快，在深度学习方面的突破可能能够解决这个问题。

不难看出，现在数据的处理实际上变成智能制造的前置条件。

再从智能制造的设备的布局来说，只有把数据传送上去，命令的指令才能传送下来。首先，现在大量的设备是不是智能的，能不能够采集数据还是个大问题。从某方面来说，有很多设备还是需要更新换代的。其次，有了数据之后，是不是能够传输到数据库中。

因为现在有很多的智能设备的控制模块和设备是分开来卖的，如果你要增加控制模块需要增加费用。有的企业为了节约成本，并没有购买设备的控制模块，虽然设备上有数据的采集和传输功能，但无法使用。

再次是网络问题。在企业、车间中需要网络，因为数据量很大、点又多，比一般的办公室规模要大得多，所以网络要成熟。

还有就是涉及一个大数据的存储和处理。一般的工厂能不能存储这么大的数据，如果数据只是监控用，只能存半年然后就丢掉，显然不能满足这个要求。

这么多数据的识别，相当于监控，只看一个时间点的数据，可以选择看一台设备。假如现在有很多设备有很多时间点的数据，这显然对技术是一个很大的考验，这是数据时代应该解决的问题。

只有数据能够存储、识别，那样才能有决策，才能智能化。所以下一个时代应该是数据时代，智能制造时代应该是再下一个时代的事情。

为什么会提出包括机器人的智能时代呢，实际上是人类在发展中比较普遍的现象，就是在目标没有明确的情况下，通常会把抽象的技术拿出来作为目标，实际上是权宜之计。在向这个目标前进的过程中，发现了一种比较适宜的技术以后，革命的方向就会改变，变成比较实际的技术，即人类是可以实现的并马上产生效果的技术，这样才会吸引大量的资金、人力、物力的投入，形成产业的一个高潮。

所以，如果把智能制造当成前面十公里的目标，可能在三公里的时候发现了数据时代的目标，会转变方向，大部分人最终进入的是数据时代，但我们的眼光依然看着十公里以外的智能制造时代。数据时代是走向智能制造时代的一个里程碑，最终还是要奔向智能时代。

作为企业，资金是有限的，经营的周期也是有限的，每天也都会有固定的成本，所以对方向的选择很重要。如果在遥远的未来，他的资金只能维持在四公里，就只能对数据技术进行投资，根本没有能力支撑他跑到十公里外的智能制造时代。

6.2 工业企业数据总体架构

一个制造企业有哪些数据，数据从哪里来，互相之间有什么关系，是一个企业信息化规划的重点。有了规划，就知道应该上什么信息系统，与现有系统相比，还有哪些缺口。

制造企业的数据来源有六层（见图 6-1），分别是：L0 层现场仪表，L1 层智能仪表；L2 层控制数据；L3 层生产数据；L4 层经营数据；L5 层财务数据。

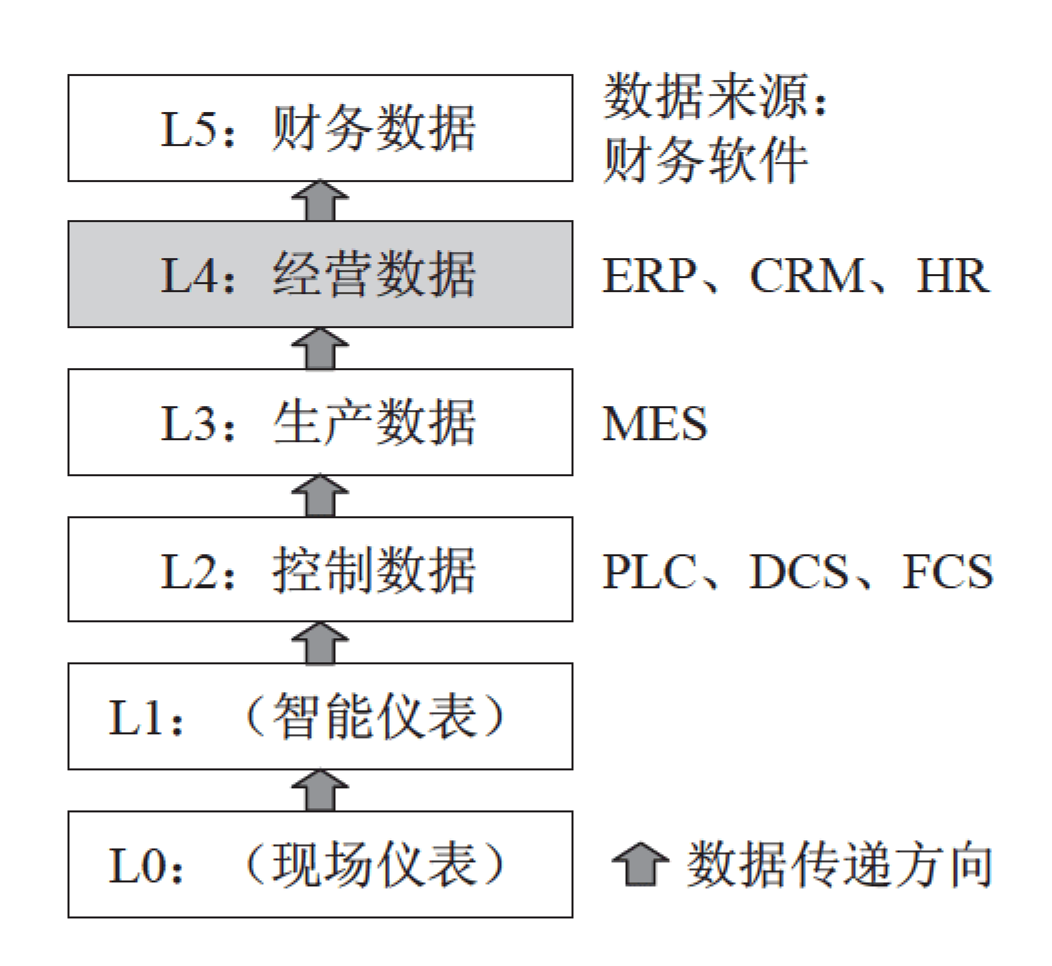


图 6-1　企业数据的总体架构

考虑到数据的归集，通过网络最低层只能读到 L2 层控制数据，所以，数据处理只有四层。

图 6-2 显示各层次数据维度和数据粒度。控制层数据来源于 PLC、DCS 或 FCS 系统，主要维度有客户、采集点、工艺参数，时间维度包括时、分、秒、亚秒，被企业车间层关注，主要关注数据采集点的平均值、最大值或最小值。

生产层数据来源于 MES 系统，主要维度有设备、物料、班次等，时间维度包括日、轮班、时、分、秒，被企业车间层和部门层关注，主要关注产量。

业务层数据来源于 ERP、CRM、DRP 等系统，主要维度有客户、存货等，时间维度包括月、周、日，被企业部门层和公司层关注，主要关注数量和

金额。

财务层数据来源于财务软件系统，主要维度有客户，时间维度包括年、季、月，被企业公司层和财务部门关注，主要关注金额。

四层之间的数据关系是：财务数据来源于经营数据，经营数据来源于生产数据，生产数据来源于控制数据。但每层的数据又不仅仅来源于下层，比如财务数据还来源于费用报销和银行，经营数据还来源于采购和仓库。

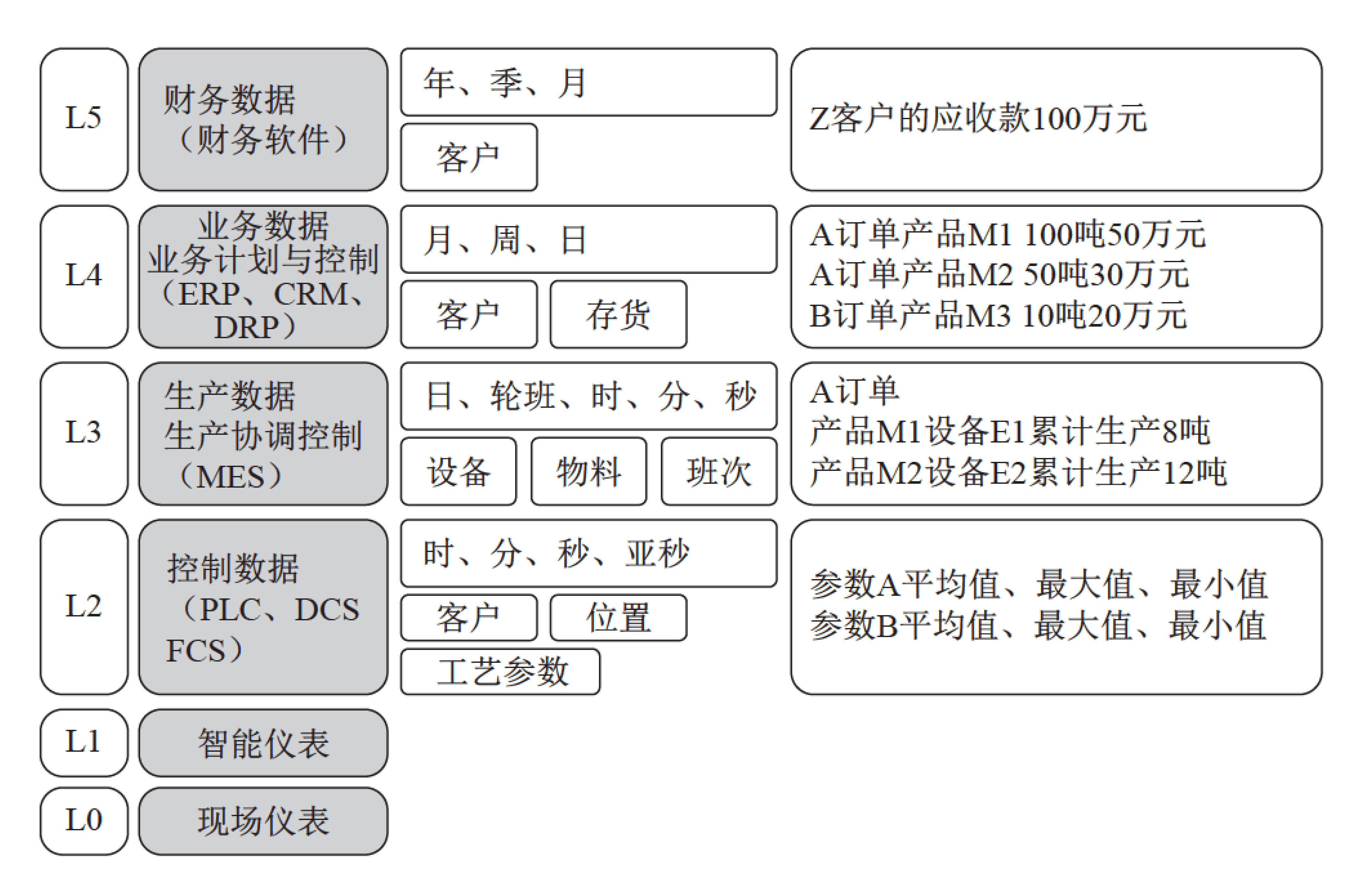


图 6-2　各层次数据维度和数据粒度

图 6-3 显示各层次数据的使用对象，公司层（决策层）关注财务和业务数据，部门层（管理层）关注业务和生产数据，车间层（执行层）关注生产和控制数据。

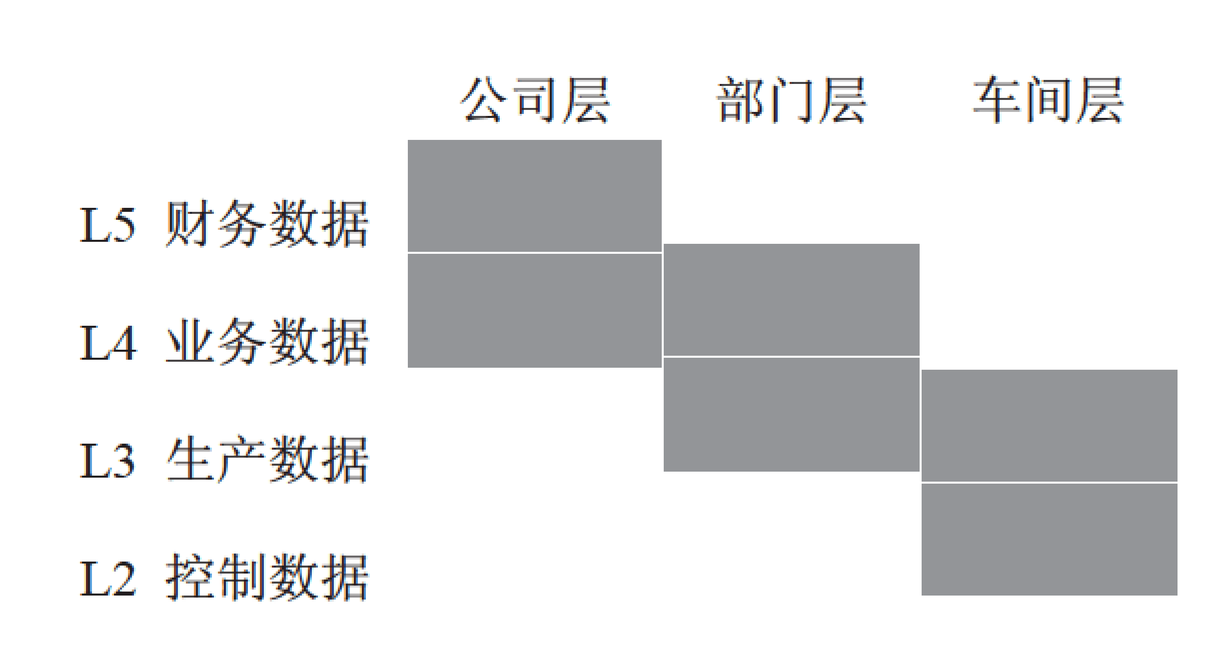


图 6-3　各层次数据的使用对象

6.3 财务数据分析

6.3.1 四个层次

财务数据来源于财务关系软件，可以从凭证中抽取数据进行分析。

财务数据分析可以分为四个层次（图 6-4）：财务凭证层、财务指标层、财务比率层和综合指标层。

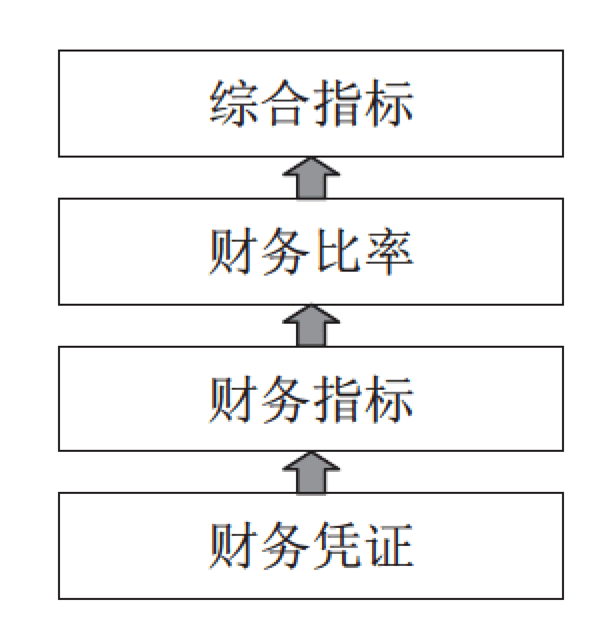


图 6-4　财务数据分析总体架构

财务凭证层直接读取财务凭证数据，可以按不同核算账簿、会计主体、科目和辅助核算查询到数据的合计及明细。

财务指标层则按资产负债表、利润表、现金流量表上的财务指标，根据组成指标的科目，用财务凭证的数据进行计算。比如，资产负债表中的流动资产由货币资金、应收票据、应收账款、预付账款、应收股利、应收利息、存货、其他应收款等多个指标构成，而货币资金可以通过科目 1001 库存现金、1002 银行存款、1012 其他货币资金用凭证中数据汇总而成。

财务比率则由财务指标计算而来。财务比率有销售利润率、营业利润率、营业成本率、期间费用率、成本费用利润率、销售费用率、管理费用率、财务费用率、经营现金净额、销售收到现金、销售现金比率、营业收入现金含量、存货周转率（次）、应收账款周转率、应收账款周转天数（天 / 次）、营业周期（天 / 次）、存货周转天数（天 / 次）等多种。

比如，存货周转率用于计算企业在一定时期内占用资金可周转的次数。

存货周转率是企业一定时期销货成本与平均存货余额的比率。用于反映存货的周转速度，即存货的流动性及存货资金占用量是否合理，促使企业在保证生产经营连续性的同时，提高资金的使用效率，增强企业的短期偿债能力。

存货周转率计算公式为销售成本 / 平均存货，销售成本直接取财务指标，平均存货需要把期初存货加期末存货除以 2。

综合指标是用一个数据评价企业的业绩好坏。综合指标主要有沃尔综合评分法和阿特曼 Z-score 模型。

6.3.2 阿特曼Z-score模型

Z-score 模型以多变量的统计方法为基础，以破产企业为样本，通过大量的实验，对企业的运行状况、破产与否进行分析、判别的系统。

Z 值用 5 个值加系数计算而来，Z 值越小，企业失败的可能性越大。

计算 Z 值的每个值来源于财务指标，对于公开上市公司，使用模型 A：

$X1$= 流动资本 / 总资产 =（流动资产 − 流动负债）/ 总资产

$X2$= 留存收益 / 总资产 =（股东权益合计 − 股本）/ 总资产

$X3$= 息税前收益 / 总资产 =（利润总额 + 财务费用）/ 总资产

$X4$= 优先股和普通股市值 / 总负债 =（股票市值 × 股票总数）/ 总负债

$X5$= 销售额 / 总资产

如果企业的 Z 值大于 2.99，企业经营处于安全区。如果 Z 值小于 1.8，则企业很有可能破产。在 1.8 和 2.99 之间，则属于灰色区。

对于非上市公司，使用模型 B，$X4$ 的计算公式不同：

$X4$= 权益账面价值 / 总负债

制造业的计算公式为：$Z=0.717X1+0.847X2+3.107X3+0.420X4+0.998X5$

如果企业的 Z 值大于 2.9，企业经营处于安全区。如果 Z 值小于 1.23，则企业很有可能破产。在 1.23 ～ 2.9，则属于灰色区。

非制造业的计算公式为：$Z=6.56X1+3.26X2+6.72X3+1.05X4$

新兴市场的计算公司为：$Z=3.25 + 6.56X1+3.26X2+6.72X3+1.05X4$

对于非制造业和新兴市场，如果企业的 Z 值大于 2.6，企业经营处于安全区。如果 Z 值小于 1.1，则企业很有可能破产。在 1.1 和 2.6 之间，则属于灰色区。

6.3.3　财务比率

财务比率是以财务报表资料为依据，将两个相关的数据进行相除而得到的比率。张燕、张樟德编著的《最实用的 120 种财务分析工具》[8] 描述了能收集到的所有财务比率，但具体使用时还需要调整，保证每个参数都可以从财务报表中读取。

财务比率按短期偿债能力、长期偿债能力、营运能力、获利能力、发展能力分为五类。

短期偿债能力中的财务比率有流动比率、现金比率、现金流量比率、现金净流量比率、现金流动负债比率、应付账款平均付账期、营运资本对总资产的比率、营运资金、营运比率、现金流动比率、经营活动的现金流量本期到期债务率、债务现金支付率、现金流量对资本支出的比率、外部融资比率。

长期偿债能力中的财务比率有资产负债率、产权比率、公积金与权益资本比率、有形净值负债率、股东权益比率、权益乘数、长期资产适合率、长期负债与固定资产比率、长期负债与营运资金的比率、债务保障比率、长期负债比率、现金负债总额比。

营运能力中的财务比率有存货周转率、存货周转天数、应收账款周转率、应收账款周转天数、营业周期、营运资本周转率、流动资产周转率、流动资产周转天数、流动资产利润率、固定资产收入率、固定资产利润率、固定资产增长率、资产周转率、总资产利润率、加速流动资产周转所增加的收入。

获利能力中的财务比率有资产净利率、流动资产利润率、流动资产营业净利率、资产现金回报率、所有者权益现金回报率、主营业务收现率、销售现金比率、营业活动收益质量比率、盈利质量比率、盈余现金保障倍数、

营运指数、全部资产现金回收率、经营现金比率、净资产收益率、资本金利润率、资本保值增值率、营业毛利率、销售利润率、销售净利率、现金流量净利率、销货收现率、成本费用利润率、全部成本费用总（净）利润率、主营业务利润率、营业利润率。

发展能力中的财务比率有资产增长率、资本积累率、销售增长率、三年销售平均增长率、利润增长率、主营业务鲜明率。

6.4 经营数据分析

经营数据也是供应链数据，包括企业的销售、采购、应收、应付和库存的数据。经营数据的来源是企业的 ERP 系统。

经营数据分析可以分为经营数据中心、销售数据分析、毛利数据分析、采购数据分析、应收款数据分析、应付款数据分析、库存数据分析七个主题。图 6-5 是程序实现的界面，左边菜单列出不同的主题，右边最上面的标签区分主题，下面横条中的标签控制同个主题的不同维度，每个统计图



图 6-5　决策支持系统的程序页面

形对应一个或多个指标，统计图形的横轴与维度相关。在统计图形上，折线图的圆点或直方图的柱子可以用鼠标单击，显示下级的数据，比如某一年或某个省。

表 6-1 显示经营数据分析中各个主题的维度，可以看出，许多主题的维度是相同的。

表 6-1 主题与维度对照表

主题 \ 维度	日期	业务流程	客户 / 供应商	部门	业务员	存货	仓库	账龄 / 库龄	入出库类型	结算方式
经营数据中心	√			√						
销售数据分析	√	√	√	√	√	√				
毛利数据分析	√	√	√	√		√				
应收款数据分析	√	√	√	√	√			√		
采购数据分析	√	√	√							
应付款数据分析	√	√	√					√		√
库存数据分析	√					√	√	√	√	

6.4.1 名词解释

（1）业务类型：按各行业中需要处理事务的不同所进行的种类划分。不同业务类型其业务的处理过程及财务收支核算的过程有差异，所以对应在系统中也会有不同的业务处理流程。一般有销售业务类型和采购业务类型，供应链管理中系统默认分为：经销、代销、直运销售、直运采购、普通采购、委托代销，等等。也可以根据企业自身情况进行自定义。

（2）信用额度：即允许客户累计欠款的最高额度。这是控制企业财务风险的一个必须要素，对不同等级的客户有不同的信用额度授权。如果超出该客户的信用额度，ERP 马上给予预警提示，并自动阻止该客户新订单或新发货，通知相关部门催款，大大降低企业应收款风险。

（3）信用账期：账期是指向客户供货后，允许客户欠款的最长时间。企业在规定时间内给予客户一定金额的信用额度，在规定时间内必须回款，这个规定时间内的周期就称为账期，额度和账期一般可以根据合作的情况进行调整。

（4）账龄分析：账龄指公司尚未收回的应收账款的时间长度，对仍在重复销售的客户而言同时也是应收账款的周转天数。通常按照各自企业合理的周转天数将其划分为四个级别，如将合理的周转天数设定为 30 天，即可分为 30 天以内、30 ～ 60 天、60 ～ 120 天及 120 天以上。

（5）库存账龄：库存账龄在 ERP 系统内，应该可以查询指定的时间点，各库存存货的库存账龄情况，即从入库起在仓库中放置了多久。与应收账款的账龄一样，存货的库存账龄越长，说明周转越慢，占压的资金也就越多。库存账龄 = ∑ (批次入库数量 × 批次入库时间 / 统计时点库存总额)。

（6）合并客户：一个集团或公司下有多个子公司，多个子公司分别作为单个客户跟企业有销售业务，但在销售数据分析中显示销售客户的排名时，子公司的排名就比较落后。但整个集团或总公司中的业务排名是比较靠前的，这时，我们可以通过设置把多个子公司合并到集团或总公司中，把单个客户汇总显示集团或总公司的排名。

6.4.2 经营数据中心

经营数据中心汇集主要的经营数据，用于监控企业的日常经营状况。每天显示前一天的年初至今合计数据，并显示当年的数据按月分布情况，按下属企业或部门的分布情况。

经营数据中心以仪表板形式，主要以企业决策层为使用对象。但如果想仔细分析某个指标，可以进入具体的主题进行 OLAP 分析，通过钻取可以看到具体的明细记录。比如看到销售收入年初至今数据不太满意，可以转到销售数据分析主题，按客户分析看到每个客户的销售数据，直至客户的每个具体订单或发货单。

图 6-6 显示最近时间主要指标的仪表板，图 6-7 显示按主要维度（时间和组织）指标，图 6-8 显示按一个指标的排名和份额。

查询维度分为日期维度和组织维度。

（1）日期维度：层次结构的级别为年、月、日，开始时间为年初，即 1 月 1 号，如果日期包括当年当月，为至当前日的前一天为止的累计值，

否则等同年度或月度数据值。

（2）组织维度：可具体分级为集团、子公司、部门。

在经营数据中心，显示以下指标：

（1）累计毛利：期初至今毛利的累计值。

（2）毛利率：累计毛利 / 累计销售收入 ×100%。

（3）累计销售数量：期初至今销售出库数量的累计值。

（4）累计销售金额：期初至今销售出库金额的累计值。

（5）应收款余额：期末核销的累计应收款金额。

（6）累计采购订单金额：期初至今采购金额的累计。

（7）累计采购入库数量：期初至今采购入库数量的累计。

（8）应付款余额：期末核销的累计应付款金额。

（9）库存数量：根据存货分类显示本期最后一天的库存数量。



图 6-6　显示最近时间主要指标的仪表板

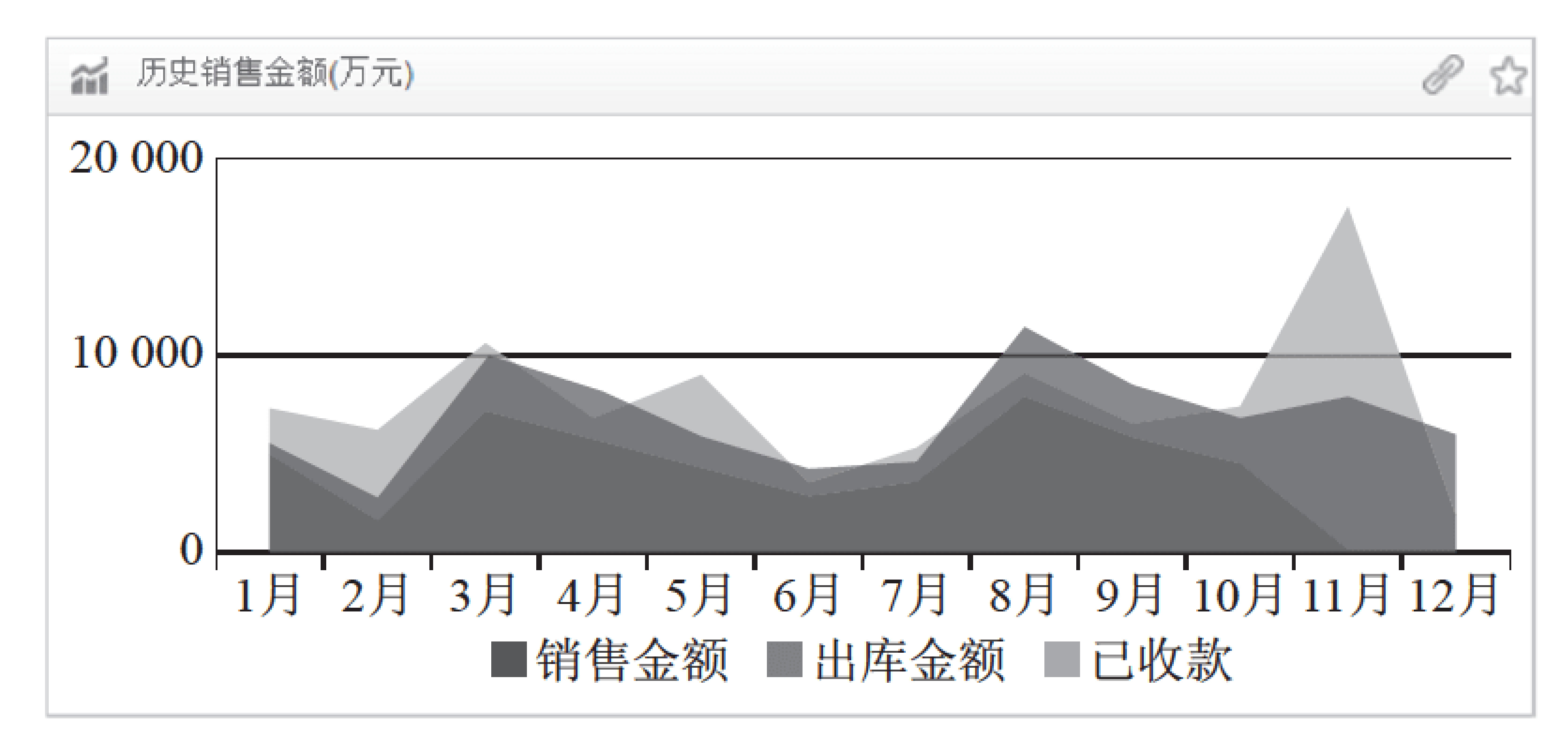
（10）历史销售金额：销售金额的时间走势。

（11）组织销售金额：各个组织销售金额的比较。

（12）客户销售金额排名 TOP10：销售出库金额最高的前十名客户。

（13）客户销售金额排名 TOP10 份额：销售出库金额最高的前十名客户的占比情况。

（14）业务人员销售金额排名 TOP10：销售出库金额最高的前十名业务员。



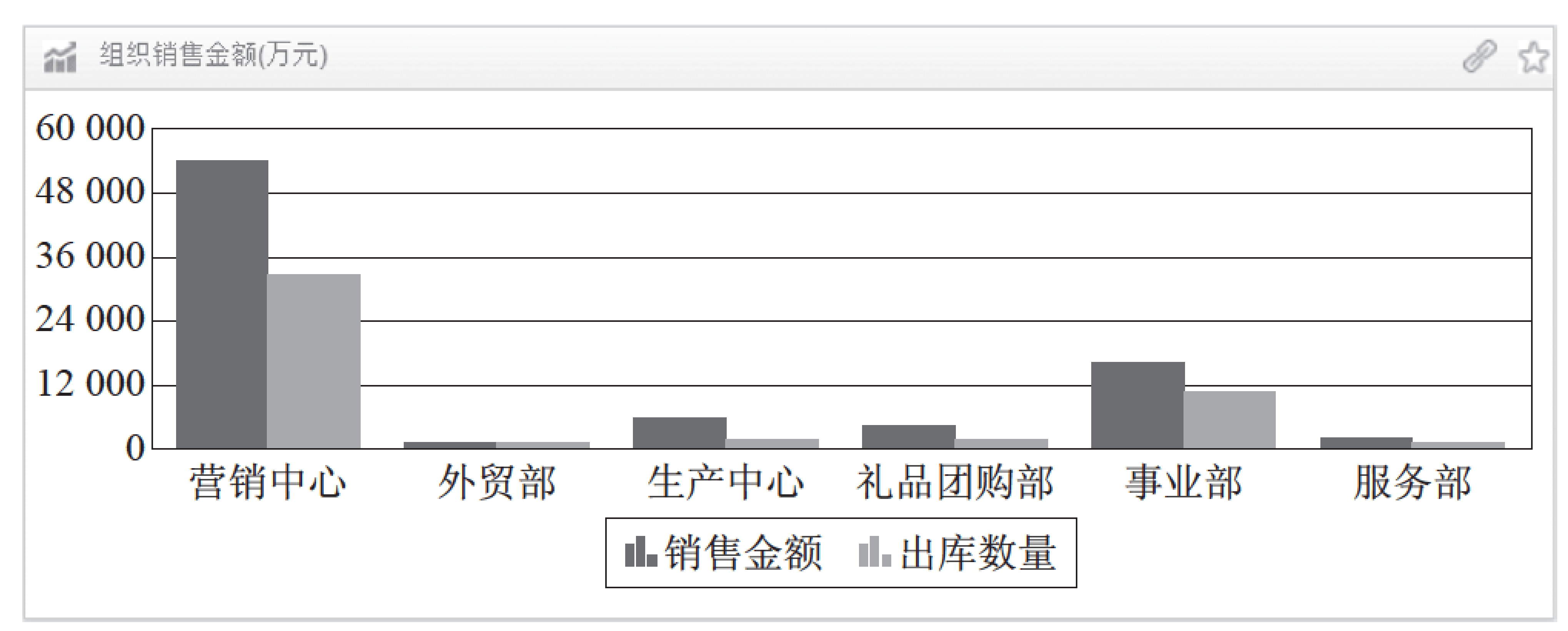


图 6-7　显示按主要维度（时间和组织）指标

（15）业务人员销售金额 TOP10 份额：销售出库金额最高的前十名业务员的占比情况。

（16）存货销售金额排名 TOP10：销售出库金额最高的前十个存货。

（17）存货销售金额排名 TOP10 份额：销售出库金额最高的前十个存货的占比情况。

（18） 销售金额增长率最高的 10 个客户：销售出库金额同比增长最快的前十名客户（去年销售出库金额取排名前 70% 的客户做比较）。

（19）销售金额下降率最高的 10 个客户：销售出库金额同比增长最慢(包括下降)的 10 名客户(去年销售出库金额取排名前 70% 的客户做比较)。

（20）销售金额增长率最高的 10 个业务员：销售出库金额同比增长最快的前 10 个业务员(去年销售出库金额取排名前 70% 的业务员做比较)。

（21）销售金额下降率最高的 10 个业务员：销售出库金额同比增长最慢（包括下降）的 10 个业务员（去年销售出库金额取排名前 70% 的业

务员做比较）。

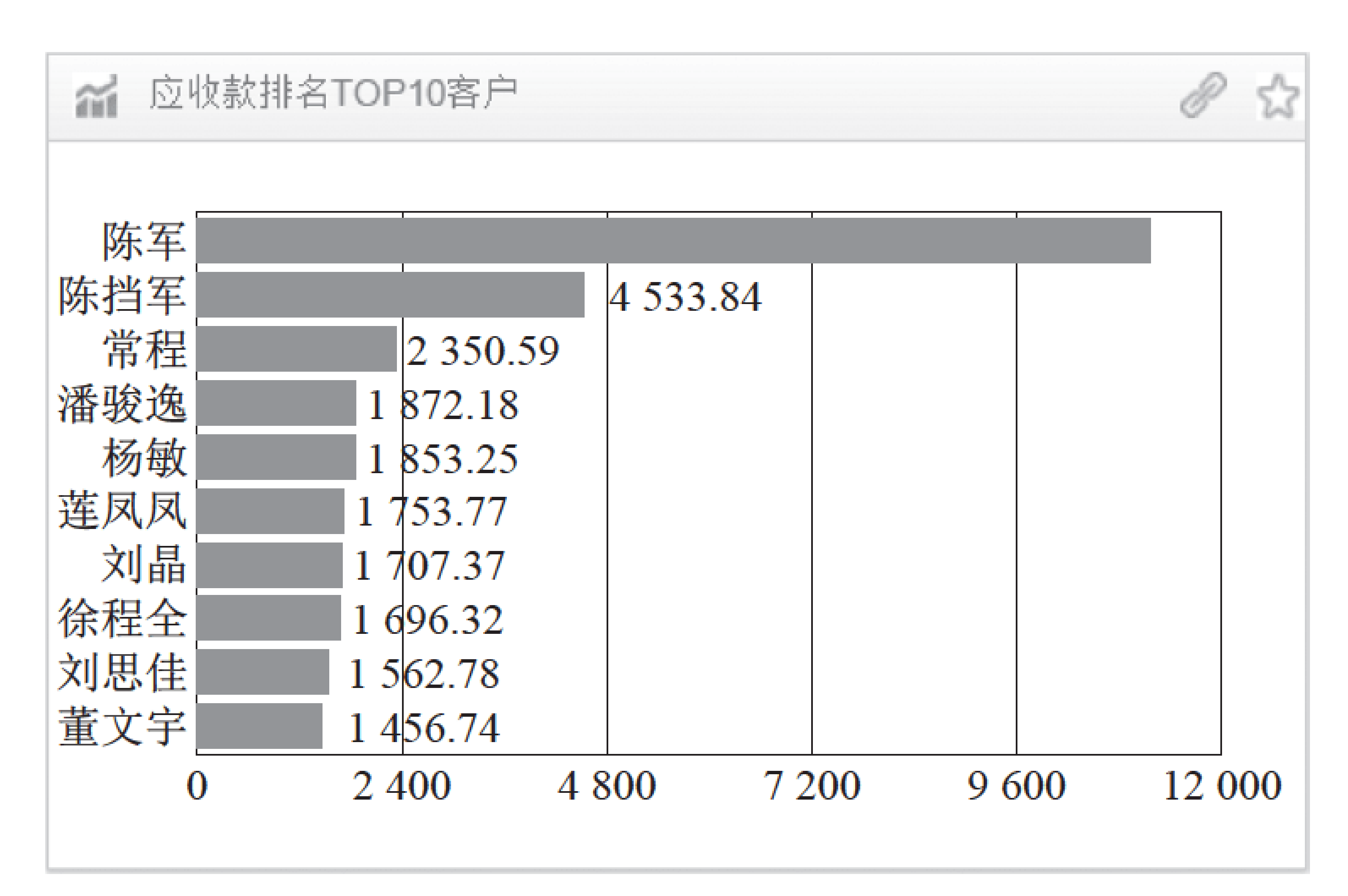
（22）销售金额增长率最高的 10 种存货：销售出库金额同比增长最快的前十种存货（去年销售出库金额取排名前 70% 的存货做比较）。

（23）销售金额下降率最高的 10 种存货：销售出库金额同比增长最慢（包括下降）的 10 名存货（去年销售出库金额取排名前 70% 的存货做比较）。

（24）应收款排名 TOP10 客户：应收款最多的前 10 名客户。

（25）应收款排名 TOP10 客户份额：应收款最多的前 10 个客户的占比情况。

（26）应收款增长率最高的 10 个客户：应收款同比增长最快的前 10 个客户（去年应收款排名前 70% 的客户做比较）。



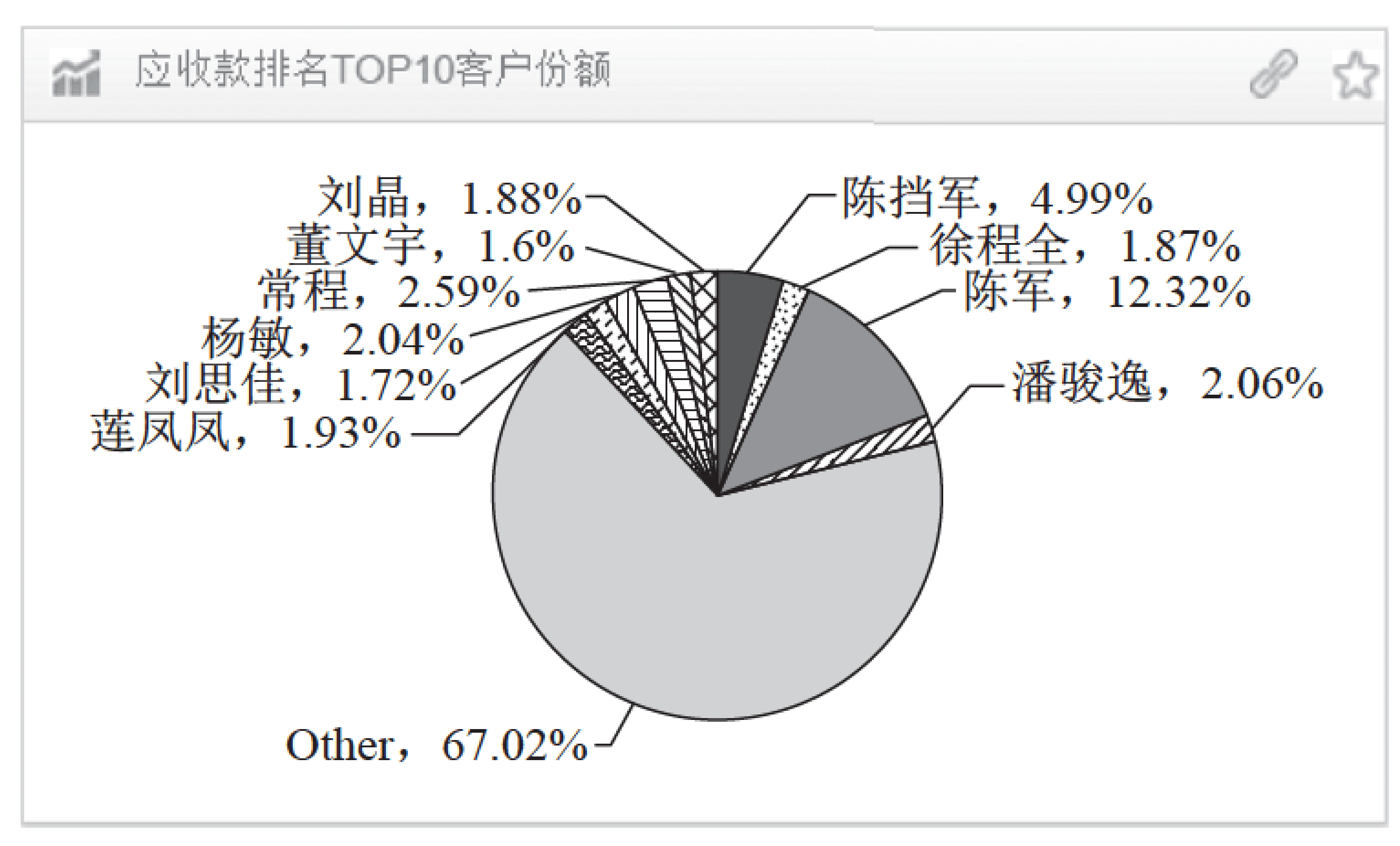


图 6-8　按一个指标的排名和份额

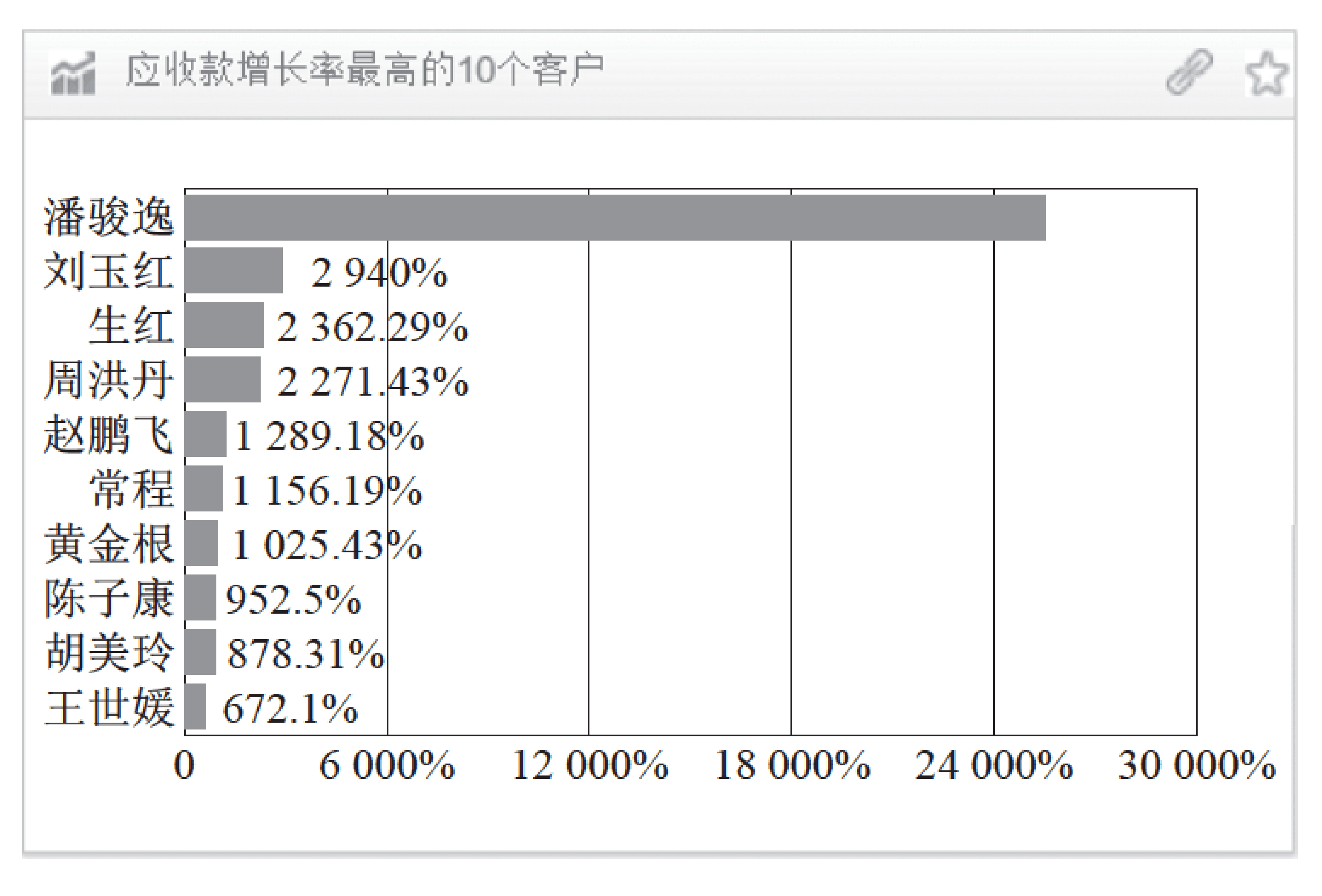


图 6-8 （续）

6.4.3 销售数据分析

销售数据分析主要对企业与销售有关的业务数据进行分析，这些业务包括销售、发货、出库、退货等。

销售数据分析的维度及相关层次结构定位为：

（1）时间：可具体分级为年、月、日。

（2）业务流程（业务类型）：可具体分级为所有、具体业务流程。

（3）地区客户：可具体分级为所有、地区、客户名称。

（4）分类客户：可具体分级为所有、分类、客户名称。

（5）销售部门：可具体分级为所有、子公司、销售部门。

（6）业务员：可具体分级为所有、子公司、部门、业务员。

（7）存货：可具体分级为所有、存货分类（可以有多层次的分类）、存货名称。

（8）存货系列：可具体分级为所有、系列名称、存货名称。

下面按销售流程，介绍销售数据分析显示的数据指标：

（1）销售合同，它在销售业务流程中是个可选流程，不是所有的销售业务中都有销售合同。与销售合同有关的指标有最高销售金额和最低销售金额，最高销售金额为销售合同中销售金额最大的一笔的单笔金额数，最

低销售金额为销售合同中销售金额最小的一笔的单笔金额数。

（2）销售订单，它是企业和客户确认要货需求的单据，一般都有销售订单和销售订单明细组成。与销售订单有关的指标有订单数量和订单金额，订单数量为销售订单模块中的订单数量，可以分为汇总数量和明细数量，汇总数量读取一个订单的多个存货的数量之和；明细数量读取每个存货的订单数量。订单金额为销售订单模块中的订单金额，可以分为汇总金额和明细金额，汇总金额读取一个订单的多个存货的金额之和，明细金额读取每个存货的订单金额。

（3）销售发货，它是企业执行销售订单，将货物发往客户的行为，销售发货单和发货单明细是客户发货的凭据，指标中的数据也来源于此。销售发货的指标有发货数量和发货金额，发货数量为销售发货单的数量，发货金额为销售发货单的金额。

（4）申请退货，主要体现的是销售发货后，客户退货的申请情况，申请退货的凭据有退货申请和退货申请单明细。申请退货的指标有申请退货数量和申请退货金额，申请退货数量为销售退货申请单中的数量，申请退货金额为销售退货申请单中的金额。

（5）销售出库，主要的单据是销售出库单，它是销售出库业务的主要凭据，在库存管理系统中用于存货出库数量、金额的核算。销售出库的指标有销售数量、销售金额、销售计划达成率、客户数量、存货品种数量，销售数量为销售出库单中的数量，销售金额为销售出库单中的金额，销售计划达成率为实际销售金额 / 计划销售金额，客户数量为在指定时间内销售出库模块（每单出库数量大于 N）的客户的个数，存货品种数量为在指定时间段内销售出库模块的存货品种（存货合计出库大于 N）个数。

（6）销售退货，主要的单据是销售出库单中出库数量为负数的单据。销售出库的指标有退货数量和退货金额，退货数量为销售实际退货的数量，退货金额为销售实际退货的金额。

大多数基本指标都可以显示同比或环比数据，称为计算指标，计算指标跟查询维度和基本指标有关，下面以客户维度的销售金额为例说明，其他维度的其他基本指标的计算相同。

（1）销售金额同比增长率：销售出库金额的同比值，月同比增长率为跟上一年本月的销售出库金额比较；日同比增长率为跟上一年这一天销售出库金额比较。

（2）销售金额环比增长率：销售出库金额的环比值，月环比为跟本年的上个月销售金额比较；日环比为跟前一天销售金额比较。

（3）销售金额 TOP10 客户：销售出库金额最高的前十个客户；二级页面中主要针对这 10 个客户做详细的分析。

（4）本期销售金额：销售出库金额最高的前十个客户的本期销售金额。

（5）同期销售金额：销售出库金额最高的前十个客户的上一年同期的销售金额。

（6）销售金额同比 %：销售出库金额最高的前十个客户的同比增长率；同比：月同比（跟上一年本月比较）、日同比（跟上一年这一天比较）；销售金额的同比查看的是本维度的同比值。

（7）TOP10 份额：销售金额最高的前十个客户关于销售金额的占比情况。

（8）TOP10 客户的时间走势：销售金额最高的前十个客户的关于销售金额的总额以及销售金额的时间走势（当期为全部时，查看各个年份的走势；当日期为年时，查看本年各个月份的走势；当日期为月时，查看本年 1 月至当月的走势；当日期为日时，查看本月 1 号到本日的走势）。

（9）同比 TOP10：销售金额同比增长最快的前十个客户（客户为同期销售金额排名的前 70% 的客户），月同比为跟上一年本月比较，日同比为跟上一年这一天比较。

（10）同比 BOTTOM10：销售金额同比增长最慢的前十个客户（客户为同期销售金额排名的前 70% 的客户），月同比为跟上一年本月比较，日同比为跟上一年这一天比较。

（11）环比 TOP10：销售金额环比增长最快的前十个客户（客户为上期销售金额排名的前 70% 的客户），月环比为跟本年的上个月比较，日环比为跟前一天比较。

（12）环比 BOTTOM10：销售金额环比增长最慢的前十个客户（客户为上期销售金额排名的前 70% 的客户），月环比为跟本年的上个月比较，

日环比为跟前一天比较。

6.4.4　毛利数据分析

毛利润是企业的运营收入之根本，只有毛利率高的企业才有可能拥有高的净利润。毛利率在一定程度上可以反映企业的持续竞争优势如何。毛利数据分析把由财务报表上一串串毛利数字，变成了直观的图形化的展示。可以直接看到这些数字产生的原因。

毛利数据分析从时间、业务流程、地区客户、分类客户、销售部门、业务员、存货维度多个方面分析货物的出库、财务成本、实际成本的数据，从而可以比较出实际成本和财务成本之间的差异，得出实际的毛利率和财务毛利率。

毛利数据分析的查询维度及相关层次结构定位为：

（1）时间：可具体分级为年、月、日。

（2）业务流程（业务类型）：可具体分级为所有、具体业务流程。

（3）地区客户：可具体分级为所有、地区、客户名称。

（4）分类客户：可具体分级为所有、分类、客户名称。

（5）销售部门：可具体分级为所有、子公司、销售部门。

（6）业务员：可具体分级为所有、子公司、部门、业务员。

（7）存货：可具体分级为所有、存货分类（可以有多层次的分类）、存货名称。

（8）存货系列：可具体分级为所有、系列名称、存货名称。

毛利数据分析显示了以下指标的数据：

（1）销售数量：销售出库单中的数量。

（2）销售无税金额（销售收入）：销售出库无税金额（出库数量 × 销售订单中的不含税单价）。

（3）销售价税合计（销售金额）：销售出库含税金额（出库数量 × 销售订单中的含税单价）。

（4）销售税额：销售价税合计 − 销售无税金额。

（5）实际成本单价：根据企业生产管理中的数据计算。

（6）财务成本：出库数量 × 财务成本单价（出库单价，无税）。

（7）实际成本：出库数量 × 实际成本单价 /1.17[出库订单、出库订单明细（无税）]。

（8）财务毛利：销售收入 – 财务成本。

（9）实际毛利：销售收入 – 实际成本。

（10）财务毛利率：财务毛利 / 销售收入 ×100%。

（11）实际毛利率：实际毛利 / 销售收入 ×100%。

以上大多数基本指标都可以显示同比或环比等计算指标，计算指标值跟查询维度和基本指标有关，具体可以获得哪些计算指标，可参考销售数据分析中的例子。

6.4.5 应收款数据分析

应收款数据分析主要展示了企业销售业务后，对客户的形成的应收款金额、已收款金额、收款余额以及账龄等数据的分析和图形的展示。能比较直观地反映出每一个客户的账款情况。从时间、地区客户、分类客户、销售部门、业务员五个条件分析应收款、已收款、账龄等数据。

应收款数据分析的维度及相关的层次结构定义为：

（1）时间：可具体分级为年、月、日。

（2）业务流程（业务类型）：可具体分级为所有、具体业务流程。

（3）地区客户：可具体分级为所有、地区、客户名称。

（4）分类客户：可具体分级为所有、分类、客户名称。

（5）销售部门：可具体分级为所有、子公司、销售部门。

（6）业务员：可具体分级为所有、子公司、部门、业务员。

（7）账龄：可具体分级为所有、账龄（0 ～ 30 天、30 ～ 60 天，60 ～ 90 天，90 天以上，也可以自定义）。

下面按应付款流程，介绍应收款数据分析显示的数据指标：

（1）销售出库，主要的单据是销售出库单，它是销售出库业务的主要

凭据，在库存管理系统中用于存货出库数量、金额的核算，也是应收款账单形成的依据。销售出库的指标有销售数量和销售金额，销售数量为销售出库单中的数量，销售金额为销售出库单中的金额。

（2）销售退货，主要的单据是销售出库单中出库数量为负数的单据。销售退货的指标有退货数量和退货金额，退货数量为销售实际退货的数量，退货金额为销售实际退货的金额。

（3）销售开票，是在销售过程中，由企业向客户开具销售发票以及发票明细的过程，它是销售收入和应收账户确认的依据。销售开票的指标有开票金额、开票价税合计、开票税额，开票金额为销售开票的无税金额，开票价税合计为销售开票的含税金额，开票税额为销售开票的税额。

（4）应收账款，是企业的往来管理系统，通过单据应收单来形成客户的应收账单。应收账款的指标有应收余额、信用额度、超期客户数、超额金额、期初余额、平均账龄、平均超账期账龄。

应收余额为期末未核销的累计应收金额，信用额度为在客户设置的额度中最后一个有效记录，超期客户数为未核销应付款中指定日期超过账期结束日的客户数（超账期金额大于等于N元），超额金额＝应收账款余额－信用额度，期初余额为上一期期末的金额。

平均账龄为单笔应收款账龄的加权平均值，而单笔应收款账龄＝每笔欠款天数（指定日－应收单日期），单笔应收账款是指与收款核销后未付款的应收账款，应收单日期是信用账期起始日。

平均超账期账龄为平均欠款天数－平均账期天数。

平均账期天数为应收账款账期天数加权平均值，应收账款账期天数为账单的到期日减去账期起效日（应收单的日期）。

（5）收款单，用来记录企业所收到的客户款项。收款单的指标有收款金额，收款金额为单据明细中的收款金额。

以上大多数基本指标都可以显示同比或环比等计算指标，计算指标值跟查询维度和基本指标有关，具体可以获得哪些计算指标，可参考销售数据分析中的例子。

6.4.6　采购数据分析

采购管理是企业供应链的重要组成部分，采购数据分析对采购的合同、订单、到货、入库、开票和采购结算的数量、金额进行对比分析和图形展示，在分析图形中还对采购的计划达成率、商品到货的合格率进行了分析，对采购合同中的最高、最低、平均价格做了比较。

采购数据分析的维度及其相关的层次结构定义如下：

（1）时间：可具体分级为年、月、日。

（2）业务流程（业务类型）：可具体分级为所有、业务流程。

（3）供应商：可具体分级为所有、地区、供应商名称。

（4）存货：可具体分级为所有、分类（多个分类）、存货名称。

（5）采购部门：可具体分级为所有、子公司、采购部门。

下面按采购业务流程，介绍采购数据分析显示的数据指标：

（1）采购请购，是采购业务处理的起点，企业内部像采购部门提出采购申请，生成采购请购单和请购单明细。该流程也是自选流程，可以选择是否需要使用。采购请购的指标有采购请购数量、最高价格、最低价格、平均价格、整体平均价格，采购请购数量为采购请购单中的货物数量，最高价格为采购请购明细单中（数量＞ N，单价＞ N）货物的最高单价，最低价格为采购请购明细单中（数量＞ N，单价＞ N）货物的最低单价，平均价格为采购购明细单中（数量＞ N，单价＞ N）的合同记录，计算货物采购加权平均单价，整体平均价格为根据采购请购单计算所有供应商的加权平均价格，所有供应商的无税采购总金额 / 采购总数量。

（2）采购订单，是整个采购业务的核心，通过采购订单可以跟踪采购的整个业务流程，主要的单据有采购订单和采购订单明细。采购订单的指标有订单数量和订单金额，订单数量为采购订单中的数量，包括总数量和明细数量，订单金额为采购订单金额，包括总金额和明细金额。

（3）采购到货，它是采购订单和采购入库的中间环节，确认对方所送货物的数量、价格和金额等信息，确认后传递到仓库，该流程也是可选流程，根据自己的业务需要选择。采购到货的指标有到货数量、到货金额、到货

合格数量、到货合格率，到货数量为采购到货单中数量，包括总数量和明细数量，到货金额为采购到货单中金额，包括总金额和明细金额，到货合格数量为采购到货后经过检验合格后的最终到货数量，到货合格率为到货合格数量 / 到货数量 ×100%。

（4）采购入库，仓库收到采购的货物，仓库将验收货物的数量，确认后入库，主要单据有采购入库单和采购入库明细单。采购入库的指标有入库数量、入库金额、采购计划达成率、供应商数、存货品种数。入库数量为采购入库单的数量，入库金额为采购入库单的金额，采购计划达成率为采购入库数量 / 采购合同数量 ×100%，供应商数为在指定时间内有交易采购入库模块的供应商的个数，存货品种数为在指定时间内采购入库模块涉及的存货品种个数。

（5）采购退货，采购到货不合格的货物，生成采购退货单。采购退货的指标有退货数量和退货金额。退货数量为采购退货单的数量，退货金额为采购退货单的金额。

（6）采购开票，是供应商开出的销售货物的发票，是销售货物的凭证。采购开票的指标有开票数量、开票税额、开票价税合计。开票数量为采购开票单的数量，开票税额为采购开票单的税额，开票价税合计为采购开票的总金额。

（7）采购结算，即采购报账，是采购核算人员跟进采购入库单和采购发票核算生成的，单据为采购结算单和结算单明细。采购结算的指标有结算数量和结算金额。结算数量为采购结算单的数量，结算金额为采购结算单的金额。

以上大多数基本指标都可以显示同比或环比等计算指标，计算指标值跟查询维度和基本指标有关，具体可以获得哪些计算指标，可参考销售数据分析中的例子。

6.4.7 应付款数据分析

应付款数据分析主要展示了企业采购业务后，对供应商的形成的应付

款金额、已付款金额、付款余额以及账龄等数据的分析和图形的展示。能比较直观地反映出每一个供应商的账款情况。

应付款数据分析从时间、供应商、存货、采购部门、结算方式五个维度分析应付款、已付款数据。维度及相关层次结构的定义为：

（1）时间：可具体分级为年、月、日。

（2）业务流程（业务类型）：可具体分级为所有、业务流程。

（3）供应商：可具体分级为所有、地区、供应商名称。

（4）采购部门：可具体分级为所有、子公司、采购部门。

（5）结算方式：可具体分级为所有、结算方式。

下面按按应付款流程，介绍应付款数据分析显示的数据指标：

（1）采购入库，仓库收到采购的货物，仓库将验收货物的数量，确认后入库，主要单据有采购入库单和采购入库明细单。采购入库的指标有采购单价、入库数量、入库金额。采购单价为采购订单含税单价，入库数量为采购入库单的数量，入库金额为采购入库单的金额。

（2）采购开票，是供应商开出的销售货物的发票，是销售货物的凭证。采购开票的指标有开票数量、开票税额、开票价税合计。开票数量为采购开票单的数量，开票税额为采购开票单的税额，开票价税合计为采购开票的总金额。

（3）采购结算，即采购报账，是采购核算人员跟进采购入库单和采购发票核算生成的，单据为采购结算单和结算单明细。采购结算的指标有结算数量、结算金额、结算单价。结算数量为采购结算单上的数量，结算金额为采购结算金额，结算单价为采购结算单对应的开票含税单价。

（4）应付账款，是企业的往来管理系统，通过单据应付单来形成客户的应付账单。应付账款的指标有应付余额、超账期应付款余额、应付余额平均账龄、平均超账期账龄、平均账期天数、提前付款的供应商数、提前付款金额、付款计划达成率。

应付余额为期末未核销的累计应付金额，超账期应付款余额为期末超过账期未核销的应付款金额。

应付余额平均账龄为应付账款账龄的加权平均值，单笔应付账款账龄 =

每笔欠款天数（指定日 － 应付单日期），应收付款是指与付款核销后未收款的应付账款。

平均超账期账龄为平均欠款天数 － 平均账期天数。

平均账期天数为应付账款账期天数加权平均值，应付账款账期天数为账单的到期日 － 账期起效日（应付单的日期），单个应付账款是指与收款核销后未付款的应付账款。

提前付款的供应商数为核销日期在到期日之前供应商数量，提前付款金额为核销日期在到期日之前的付款合计金额，付款计划达成率为付款计划达成率＝实际付款额 / 应付款额，应付款额＝查询期期初的应付余额＋查询期的应付金额 － 查询期期末未到期金额。

（5）付款单，用来记录向供应商付款的款项。付款单的指标有已付款金额，已付款金额为单据明细中的付款金额。

6.4.8　库存数据分析

库存管理系统是一个企业、单位不可缺少的部分，它的内容对于企业的决策者和管理者来说都是至关重要的。库存数据分析主要从时间、仓库、入出库方式、存货、存货系列、库龄等多方面查询度分析库存数据和图标的展示。通过分析系统可以让管理部门直观地全方位地看到库存的信息，这样既可以保证日常的生产不至于因为原材料不足而导致停产，确保生产顺利进行，也可以使企业不会因原材料的库存数量过多而积压企业的流动资金，从而提高企业的经济效益。

库存数据分析从以下维度及相关的层次结构进行分析。

（1）时间：可具体分级为年、月、日。

（2）仓库：可具体分级为所有、仓库名称。

（3）出入库类型：可具体分级为所有、出入库类型。

（4）库龄：可具体分级为所有、库龄。

（5）存货：可具体分级为所有、分类（多层次分类）、存货名称。

（6）存货系列：可具体分级为所有、系列名称、存货名称。

下面按入出库流程，介绍入出库有关的数据指标：

（1）入库业务：仓库收到采购或生产的货物，保管员将验收后的数量等信息确认后入库，单据主要包括采购入库单、产成品入库单和其他入库单。入库业务的指标有入库数量、入库金额、入库平均单价。入库数量为货物入库单上的数量，入库金额为货物入库单上的金额，入库平均单价为加权平均单价，入库总金额 / 入库总数量。

（2）出库业务：其指标有出库数量、出库金额、出库平均单价。出库数量为货物出库单上的出库数量，出库金额为货物出库单上的出库金额，出库平均单价为加权平均单价，出库总金额 / 出库总数量。

库存数据分析通过复杂的处理，生成并显示了以下指标。

（1）期末库存数：本期最后一天的库存数量。

（2）库存平均数量：每天库存数量之和 / 天数。

（3）平均库存金额：每天库存金额之和 / 天数。

（4）库存余额：本期最后一天的库存金额。

（5）财务库存毛利：销售标准价 × 库存数量 – 财务库存成本。

（6）实际库存毛利：销售标准价 × 库存数量 – 实际库存成本。

（7）当天标准售价：最新销售价格 – 销售调价单、销售调价明细单取当天的零售价，如果取不到该价格，就读取最近一次的销售价格（离选定日期最近的一次）。

（8）进货价（入库价）：最近一次入库单价。

（9）财务库存潜亏：$\sum$（进货价 × 库存数量）– 财务库存成本。

（10）实际库存潜亏：$\sum$（进货价 × 库存数量）– 实际库存成本。

（11）平均日销售数量：指定日期前 15 天的销售总数量 /15。

（12）可销售天数：库存数量 / 平均日销售数量（最近 15 天平均出库数量）。

（13）平均库龄：根据入库日期倒排序物料，用库存总数去对比入库数量，找到符合条件的多条记录（如图 6-9 所示，符合条件的 3 条）。

序号	入库日期	入库数量	入库单价	库存（2015-12-29）
1	12月25日	50	3.5	100
2	12月23日	40	2.8	
3	12月18日	20（只有10个数量符合）	4.2	
4	12月5日	50	3.0	
5	11月28日	20	4.0	
6	11月3日	30		

图 6-9　平均库龄计算

分别计算这 3 条记录中的入库单价、数量、天数（当天－入库日期），用（每笔的库存天数 × 每笔的数量）之和 / 总库存数量，得到库龄：（4×50＋6×40＋11×10）/100 ＝ 5.5（天）。

（14）财务库存成本：每笔财务入库单价×每笔入库数量之和：50×3.5＋40×2.8＋10×4.2 ＝ 329（万元）。

（15）实际库存成本：每笔实际入库单价 × 每笔入库数量之和：50×3.5 ＋ 40×2.8＋10×4.2 ＝ 329（万元）。

（16）期初库存数：上一期的期末库存数。

以上大多数基本指标都可以显示同比或环比等计算指标，计算指标值跟查询维度和基本指标有关，具体可以获得哪些计算指标，可参考销售数据分析中的例子。

6.5　与上市公司外部数据比较

在企业决策中可以利用的数据有两部分：一部分是内部数据，另一部分是外部数据。内部数据一般是来自以 ERP 为核心的企业信息系统，包括财务数据、经营数据、生产数据和一些控制数据，分别来自财务软件、ERP 软件、MES 软件和底层的 DCS 等控制系统的软件。这些内部数据相对而言比较详细、准确，也没有数据安全、使用权限问题，而且粒度比较细。对内部数据可以采取纵向比较，即按照时间进行比较分析。做数据分析的

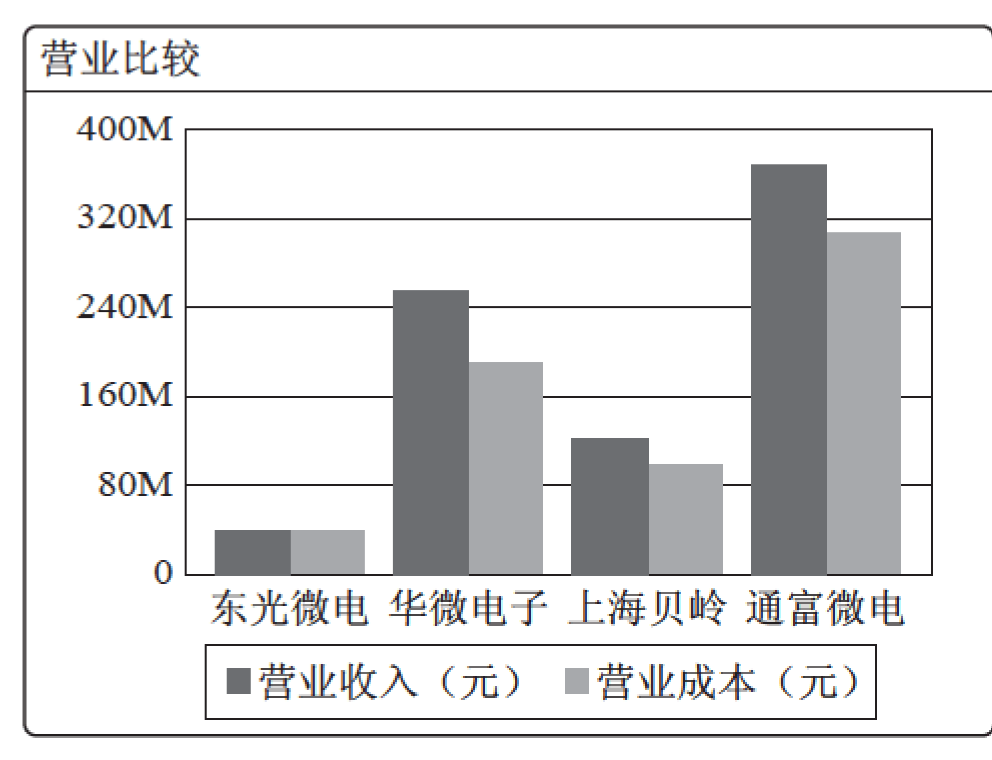
另一种方式叫横向比较，就是和你的同行进行比较。同行比较最主要的问题就是缺少同行的数据，好在如果同行已经上市，它根据证监会的要求，需要定期公布经营数据。公布时间每个季度一次，一年四次，包括第一季度报告、半年报告、第三季度报告和年度报告，这里面公开了很多的经营数据。

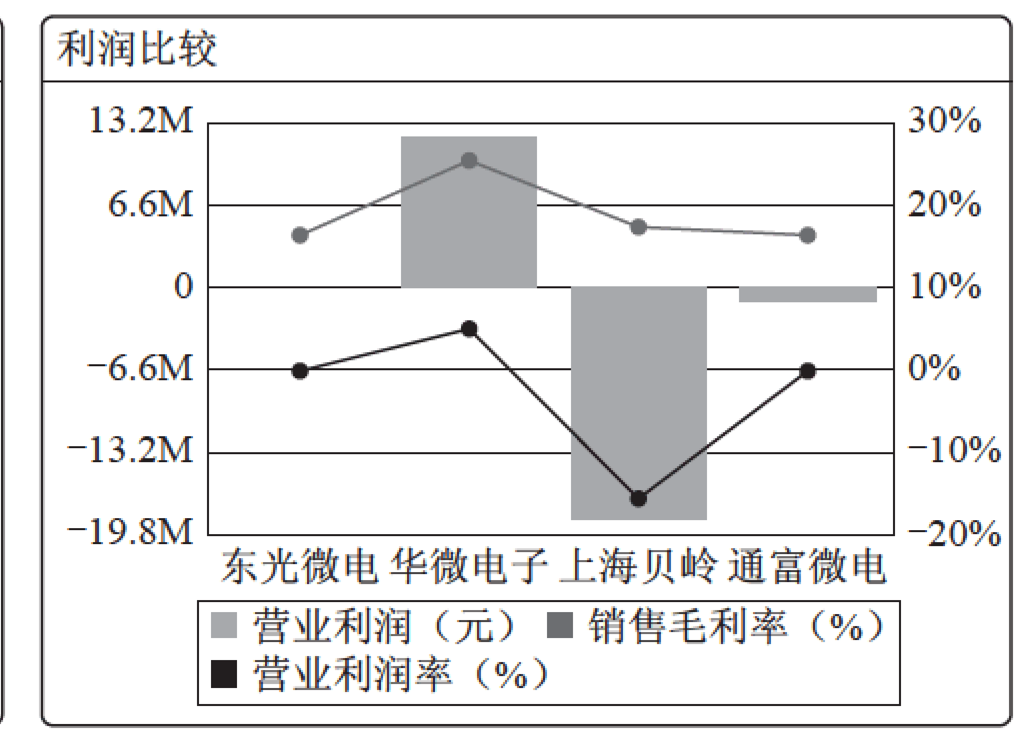
现在炒股的人一般都关注股价的数据，对经营数据不太重视，实际上从价值投资角度来说，应该重视经营数据。企业分析自己在同行中的位置时，不能使用股价数据，只能用经营数据。这些外部数据以季度为单位，没有企业内部数据的粒度高，所以进行对照分析的内部数据也只需要季度的财务数据。

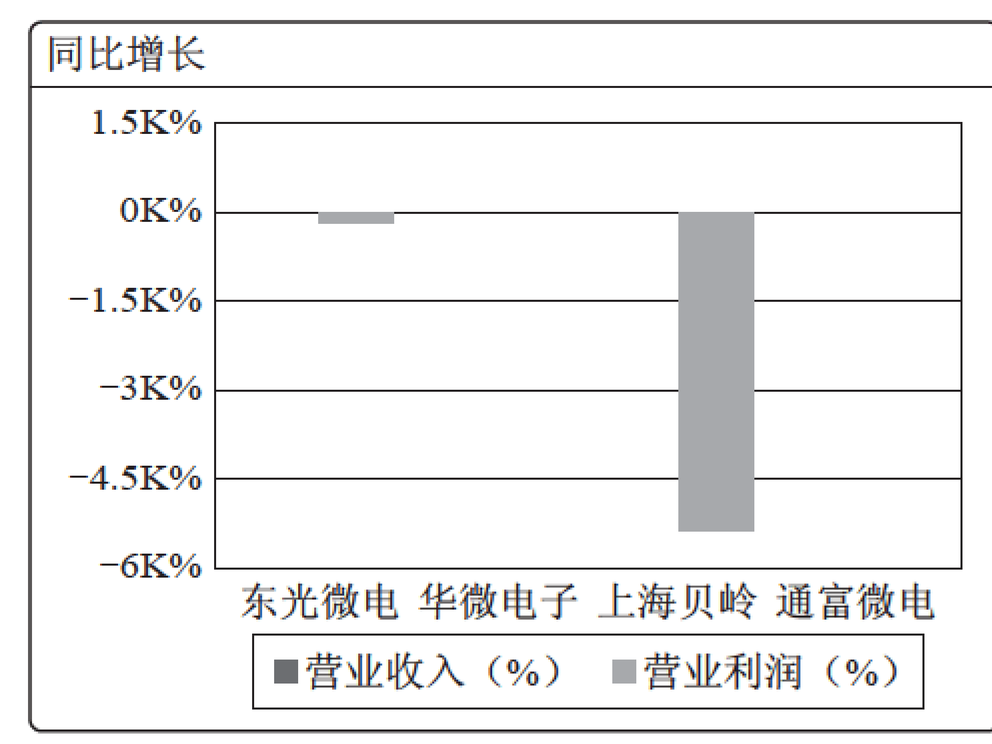
经营数据的横向比较有什么价值呢？横向比较可以发现自身不足，知道行业的增长速度，看看自己是否赶上或超过行业的平均发展水平。有时虽然自身企业的增长情况不错，比如说与去年相比增长 10%，如果这个行业平均增长 20%，那么 10% 的增长速度是远远不够的，有可能表明现在市场时机比较好，应该抓住这个机会进行发展，若是不能抓住这个机遇，过段时间你的规模就会明显地落在同行之后，会被行业所淘汰。有了横向比较才能做到心中有数。还可以对关键指标进行同行比较，比如人均销售额、人均利润等指标，看看同行的企业中本企业的劳动生产率情况如何，根据行业先进水平对企业进行调整，对自己产品的价格在市场上是否有竞争力做到心中有数，若是发现差距则可以找到差距所在，比如说人员总数太多或者开发人员占比太少等原因，及时进行调整。

上市公司定期报告数据分为合并报表和母公司报表，作为同行，主要比较母公司报表数据。

图 6-10 显示四个同行上市公司几个数据指标的对比分析统计图形。







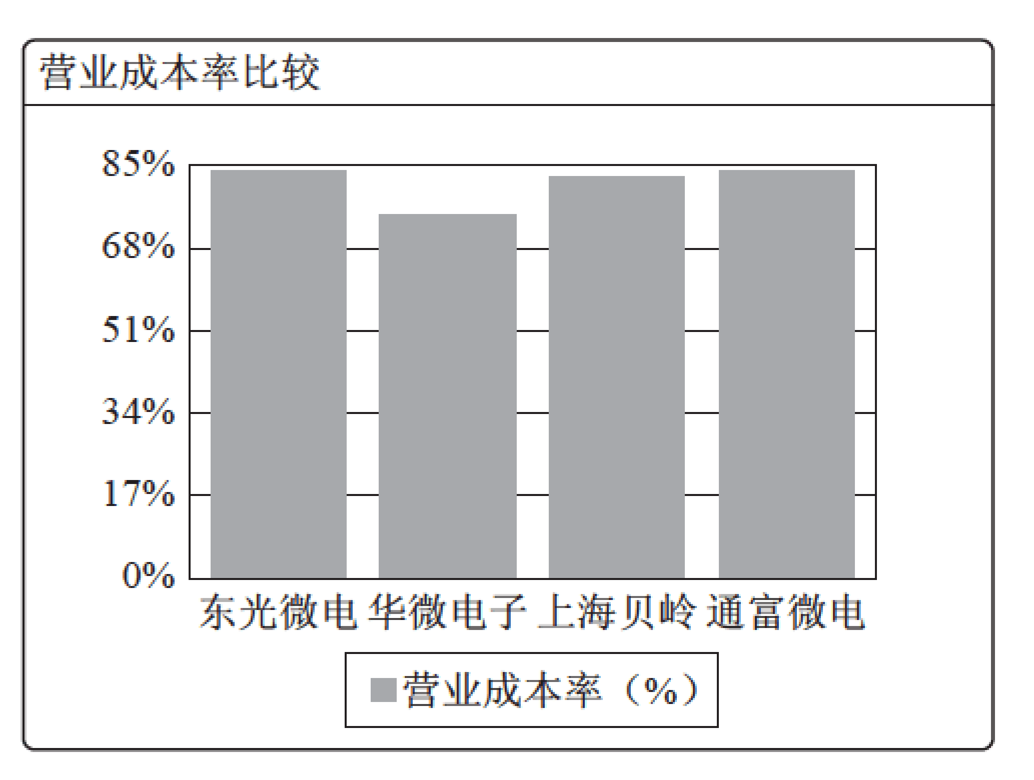


图 6-10　上市公司数据对比分析

6.6　控制数据分析

6.6.1　从工业大数据中找到故障

在李杰的《工业大数据》[12] 一书中，提出了一些非常有价值的理念：现在工业企业中处理的问题都是已经发生的问题，已经发生的故障。怎样解决没有发生的故障呢？书中举了例子，通过研究，发现可以提前十多天发现生产流水线上一个工业机器人故障。至于如何发现的，书中没有给出具体说明。

实际上，每个故障的发生，都是一些小的错误积累到一定程度的结果。通过分析设备状态多种数据，只要找到开始的一些数据变化，就可以预测

故障的发生，甚至知道故障发生的时间。

如果一个生产车间只有几十台设备，就可以逐个观察它的数据，或者通过一些统计软件对统计数据进行分析，得出结论。但如果一个企业有几千、几万台设备，怎么发现其中一台的故障呢？

如何在这么多数据中，几千台设备的运行数据中找到这个问题数据，这是一个大数据问题，李杰的书中并没有给出具体方法。

工业设备的数据正常情况下变化不大，数值比较稳定。作为一种采样数据，采样间隔如果是 10 秒，每个采集点每过 10 秒钟就会采集一个数据。设备在正常运转时，这种数据都会围绕一个基准值波动，也就是说大量的时间点采集和大量的设备传出的数据都是没有价值的。如果直接处理这些数据，会耗费大量的时间和存储空间。

可以为同类设备同种采集点设置一个基准值和正负误差范围，仅仅把与基准值有偏差的数据进行抽取，通过预处理后，然后把这个值放到数据仓库中去，最后通过建模，研究分析汇总数据。

这个模型维度主要考虑时间、设备和设备类型，设备按工厂可以分为总公司、分公司、工厂、车间、车间区域、设备号。

把设备数据按维度进行汇总，最终决策者看到的数据是所有设备汇总的数据，比如几千台设备汇总后得到的一个平均值。平均值可能是一天的平均值，也可能是一个月的平均值，甚至是更长的一年的平均值。一般只需要看一天或一个月的平均值。对工业设备来说，比较关注一天的平均值，这个平均值既可能是在时间维度的平均值，也可能是在设备维度的平均值。

要发现故障，主要需要关注平均值环比或者同比变化。因为不可能所有设备一起发生故障，所以要找的是环比或同比的排名，这个排名可以从几千台设备中挑出变化最大的前十个来。通过排名关注看设备状态有没有发生变化。

正常情况，环比应该是零或接近零，比如是百分之 0.001，这些值的排名都靠后。如果有一台设备发生故障，监测值的总和、平均值、同比或环比值肯定会发生变化，但因为数据太多，会被淹没掉，这时就需要看排名。从排名中看到，平时最高的排名环比值可能是 0.01%，现在忽然增加

到 0.1% 或者 1%。看到这种情况，就可以马上调出这台设备的历史数据，观察它在时间维度上的走势。如果以前没有类似故障的数据模型，起码应该对设备异常引起警觉，加强观察或马上停机检修。在这台设备发生故障或检修发现问题后，可以把这台设备的异动变化数据建一个模型保存起来，或者把时间的变化做一个截图保存起来，以后再看到发生类似变化的话，就可以对照这个时间走势曲线图，大概知道这个信号出现以后，比如说在 5 天后就可能发生类似的故障。

由于大型企业设备不但数量多，类型也多，每个设备又有很多数据采集点，不同采集点采集的数据类型，有的是状态值，有的是连续数值，计量单位也不同。显然，要同时监控与分析这么多数据是比较困难的。

基本的方法是对于同一类采集点，比如温度，在基准值相同的情况下，通过平均值、与基准值的偏差进行分析，这样不管设备和单台设备的采集点数量，都可以找到故障点。在检测点类型及基准值分级不多的情况下，这种方法比较简单，但如果采集点类型众多，需要分析的数值就比较多。

另一种方法是为每类设备设计一个指数，同时反映所有采集点的数据变化。把同一台设备上不同采集点的数据通过加权求和计算为一个指数，这个指数在每次采样时都保存在数据库中。计算指数的数据不能直接用采集数据，这个数据最好是跟量纲无关，如果用数值的偏差比例，这个比例值就是一个无量纲的数值。加权的权重值根据检测点数据的重要性确定。

把所有设备的指数进行统计，取它的平均值，平时监测所有设备的平均的指数值，观察平均值的变化和排名，发现有异常，再进入不同设备和采集点进行详细分析。

发现检测点的问题后，查看检测点的历史数据，历史数据会反映一种上升或下降的趋势，结合实际故障或检修结果，就可以对故障进行分类，找出故障发生的规律。

6.6.2 从检测大数据中发现质量问题

制造企业在生产过程中自动检测装置会产生大量检测数据，大型工业

企业分为子公司、车间、机台，不同子公司或车间生产不同产品，不同产品或同个产品的不同生产过程有不同的检测要求，产生计量单位和数量级不同的检测数据。

检测大数据分析的目的是从这些检测数据中找到异常情况，找到异常的检查点，追踪到子公司、车间、机台，或班组、员工，或原材料供应商。

质量数据的异常可用标准差来衡量，但如果检测数据的量纲不同，应该用变异系数。考虑到多个产品、多个检测点的实际情况，质量大数据分析要用变异系数分析。

具体流程是：通过鼠标交互，按任意维度对图形化的检测数据的变异系数进行分析，发现质量问题（变异系数值太大），再通过逐级钻取对问题进行分析，发现问题所在（比如与设备、员工还是原材料供应商有关）。

为保证不但满足目前需求，而且满足未来需求，因此不仅考虑当前关注的、敏感的指标，而是对所有检测数据均要可以进行分析。

每个检测指标提供样本个数、平均值、最大值、最小值、标准差、变异系数等统计指标。根据不同的分析条件，统计指标有不同的值。

提供基于日期、时间、生产机台、生产人员、原材料供应商维度的查询，条件可以任意组合。提供数据比较，比如不同生产机台、不同生产人员、不同原材料供应商同维度比较；不同时间同比、环比比较。

可以进行数据钻取，比如日期维度可以从年度变异系数，到某月、某日的变异系数；生产机台维度可以从车间变异系数，到某区域，到某机台变异系数。

下面介绍两个典型使用场景，一个是全面质量检查，另一个是日常质量数据检测。

全面质量检查分别按不同维度查看变异系数数据，掌握合理的变异系数值，作为质量控制的目标。分析的目的是在不同维度找到偏高或偏低的数值，比如在生产机台维度，发现车间 1 的值偏高，点击车间钻取到区域比较，发现区域B值偏高，再钻取区域B到生产机台比较，发现A机台偏高。将生产机台维度锁定为 A 机台，进入日期维度，查看 A 机台的历史数据，从年到月，再到日，可能发现 A 机台在 10 天前数据开始偏离正常值。

根据以上分析结果，可以现场安排检查 A 机台，是否出现故障。

日常数据监测可以查看昨天变异系数的同比、环比值，如同比、环比接近 0，证明质量稳定。如果发现环比值比较大，表明昨天质量出现问题，需要找到原因。

查看不同维度的环比值，找到质量问题原因。比如在原材料供应商维度，发现 A 公司供货的产品变异系数高于其他供应商。在原材料供应维度锁定 A 公司，进入日期维度，查看利用 A 公司原材料生产产品的历史数据，可能最后发现该公司原材料昨天刚投入使用。

安排仓库和采购部门对 A 公司原材料质量进行检查，可能因此及时发现一个不合格供应商或一批不合格原材料，避免继续生产导致的更大损失。





第 7 章

设计案例

7.1　政府房产数据分析

房地产在中国经济中占有非常重要的地位，国家的很多宏观财政金融决策都与房地产的投资密切相关。地方经济同样受房地产投资和销售的影响巨大，地方政府的经济决策都依赖对房地产数据的掌握。各地住房保障和房产管理局拥有房地产交易的详细数据，可以据此准确了解城市房屋的存量规模和交易情况，无论是政府政策制定、人大的议案提案、局领导的日常工作、市民的购房售房决策，都希望能够方便、直观地了解这些数据。

该案例从三个主题分析数据，分别为：监控中心、预售分析和成交分析。

7.1.1　监控中心

监控中心提供对最新数据和主要数据指标的监控和分析。

监控中心只与时间维度有关，与其他维度无关，可以查看任意日期的一些综合指标。监控中心从概要和时间两个维度分析各个指标，相应分为概要和时间维度两个标签页。

指标包括基本指标和计算指标。基本指标有平均预售基价（元 /m^2）、预售总建筑面积（m^2）、销售面积（m^2）、销售套数。

计算指标有：

（1）年初至今预售总建筑面积（m^2）：本年 1 月 1 号到所选日期的累计预售面积。

（2）年初至今销售面积（m^2）：本年 1 月 1 号到所选日期的累计销售面积。

（3）年初至今销售套数：本年 1 月 1 号到所选日期的累计销售套数。

（4）项目预售基价 TOP10：预售基价最高的 10 个项目。

（5）项目预售总建筑面积 TOP10：预售总建筑面积最大的 10 个项目。

（6）项目销售面积 TOP10：销售面积最大的 10 个项目。

（7）项目销售面积 TOP10 份额：销售面积最大的 10 个项目的百分比。

（8）项目销售套数 TOP10：销售套数最多的 10 个项目。

（9）项目销售套数 TOP10 份额：销售套数最多的 10 个项目的百分比。

概要标签页主要分析的是各个指标的合计数据以及相关指标的排名情况，用于平时监控预售数据和成交数据。概要页面分析的相关指标有平均预售基价、预售总建筑面积、销售面积、年初至今预售总建筑面积、年初至今销售面积、销售套数、年初至今销售套数，以上指标使用仪表盘表示，如图 7-1 所示，指针位置为本期销售面积值，红色与绿色交界的位置为销售面积的上期值。

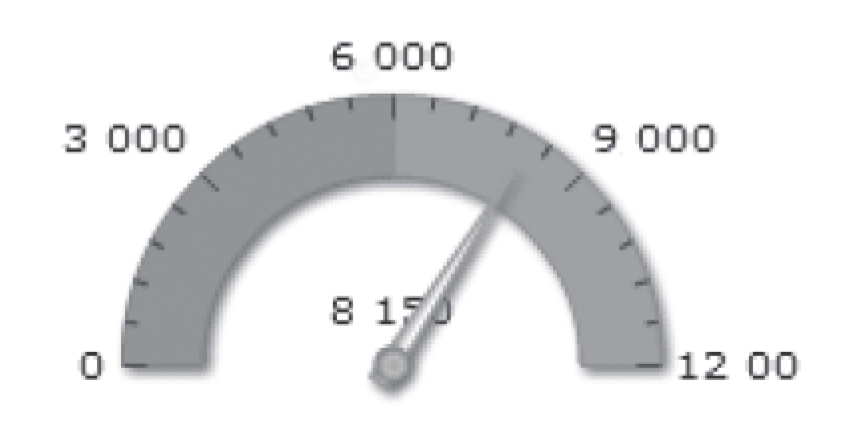


图 7-1　典型仪表盘

概要中还统计了项目预售基价 TOP10 排名，项目预售总建筑面积 TOP10 排名，项目销售面积 TOP10 排名以及 TOP10 的份额，项目销售套数 TOP10 排名以及 TOP10 的份额，通过条形图分析销售套数最高的前 10 个项目，通过饼图分析销售套数最高的前 10 个项目的占比情况。

时间维度标签页可以查看各个年份、月份及每天的指标的走势情况。通过折线图分析平均预售基价，预售总建筑面积、销售面积，年初至今预售总建筑面积、年初至今销售面积，销售套数、年初至今销售套数的走势情况。将预售总建筑面积和销售面积放在一张图形上面，既可以分析预售总建筑面积和销售面积的走势情况，也体现了预售面积与已售面积的对比关系。

7.1.2　预售数据分析

预售数据分析可以从预售时间、项目、用途三个维度进行分析，并相应地分为时间、项目、用途三个维度标签页和一个明细标签页。

预售时间维度按年、月、日三个级别的层次分析。项目维度按板块、项目名称两个级别的层次分析，一线城市需要加一个级别——区。物业类型维度只有一个级别，分为住宅或商业两项。

预售分析的指标分为基础指标和计算指标。基础指标为：预售基价（元/m^2）、预售总建筑面积（m^2）、销售面积（m^2）、项目个数、预售证号个数。

计算指标为：

（1）同比增长率：指标对于上一年数据的增长率。

（2）环比增长率：指标对于本年上一个月数据的增长率。

（3）TOP10：一个维度下最后一个层次的某个指标的前 10 名。

（4）BOTTOM10：一个维度下最后一个层次的某个指标的后 10 名。

（5）TOP10 份额：对于某个指标 TOP10 及其他的占比情况。

（6）年初至今：本年 1 月 1 号到所选日期的累计值。

（7）迄今为止：数据产生日到所选日期的累计值。

时间维度标签页可以查看任意板块、任意项目，各个年份、月份及每天的各个指标的走势情况，通过折线图可以分析平均预售基价、预售总建筑面积、销售面积、项目个数、预售证号个数等指标的走势情况。将预售总建筑面积和销售面积放在一张图形上，既可以分析预售总建筑面积和销售面积的走势情况也体现了预售面积与已售面积的对比关系。

项目维度标签页可以查询任意时间的各个板块、各个项目的相关指标的对比情况，通过直方图分析平均预售基价、预售总建筑面积、销售面积、项目个数、预售证号个数等指标的对比情况。将迄今为止预售总建筑面积和迄今为止总销售面积放在一张图形上对比，更加明显地体现预售总建筑面积和总销售面积的对比关系。

通过钻取功能，可以查询任意板块下所有项目的对比情况，比如钻取

主城区，可以查看主城区下各个项目迄今为止总预售建筑面积和迄今为止总销售面积的对比情况。

对于每个典型指标，可以查看该指标的更加详细的分析数据。以项目维度的平均预售基价指标为例，可以展示平均预售基价的本期值、同期值、同比增长率、环比增长率、平均预售基价最高的前 10 个项目、前 10 个项目的时间走势、平均预售基价同比增长最快的前 10 个项目及同比增长最慢的前 10 个项目、平均预售基价环比增长最快的前 10 个项目及环比增长最慢的前 10 个项目。

明细标签页以表格形式展示了有关预售信息的明细数据，包括预售证号、项目名称、开发企业、所在板块、房屋坐落地址、预售时间、预计竣工时间、房屋用途性质、预售总建筑面积、车库建筑面积、预售基价的详细信息。

7.1.3　成交数据分析

成交分析主要从时间、项目、面积、户型、物业类型五个维度进行分析，分别对应一个标签页，另有一个明细标签页。

时间维度按年、月、日三个级别的层次分析。项目维度按板块、项目名称两个级别的层次分析，一线城市需要加一个级别——区。面积和户型只有一个级别。物业类型维度只有一个级别，分为住宅或商业两项。

成交分析的指标分为基础指标和计算指标。基础指标为：销售套数、销售面积。

计算指标为：

（1）同比增长率：指标对于上一年数据的增长率。

（2）环比增长率：指标对于本年上一个月数据的增长率。

（3）TOP10：一个维度下最后一个层次的某个指标的前 10 名。

（4）BOTTOM10：一个维度下最后一个层次的某个指标的后 10 名。

（5）TOP10 份额：对于某个指标 TOP10 及其他的占比情况。

时间维度标签页可以查看任意板块、任意项目名称、任意面积分类、任意户型的各个年份、月份及每天的指标的走势情况。通过折线图分析销

售套数、销售面积的走势情况，如图 7-2 所示。

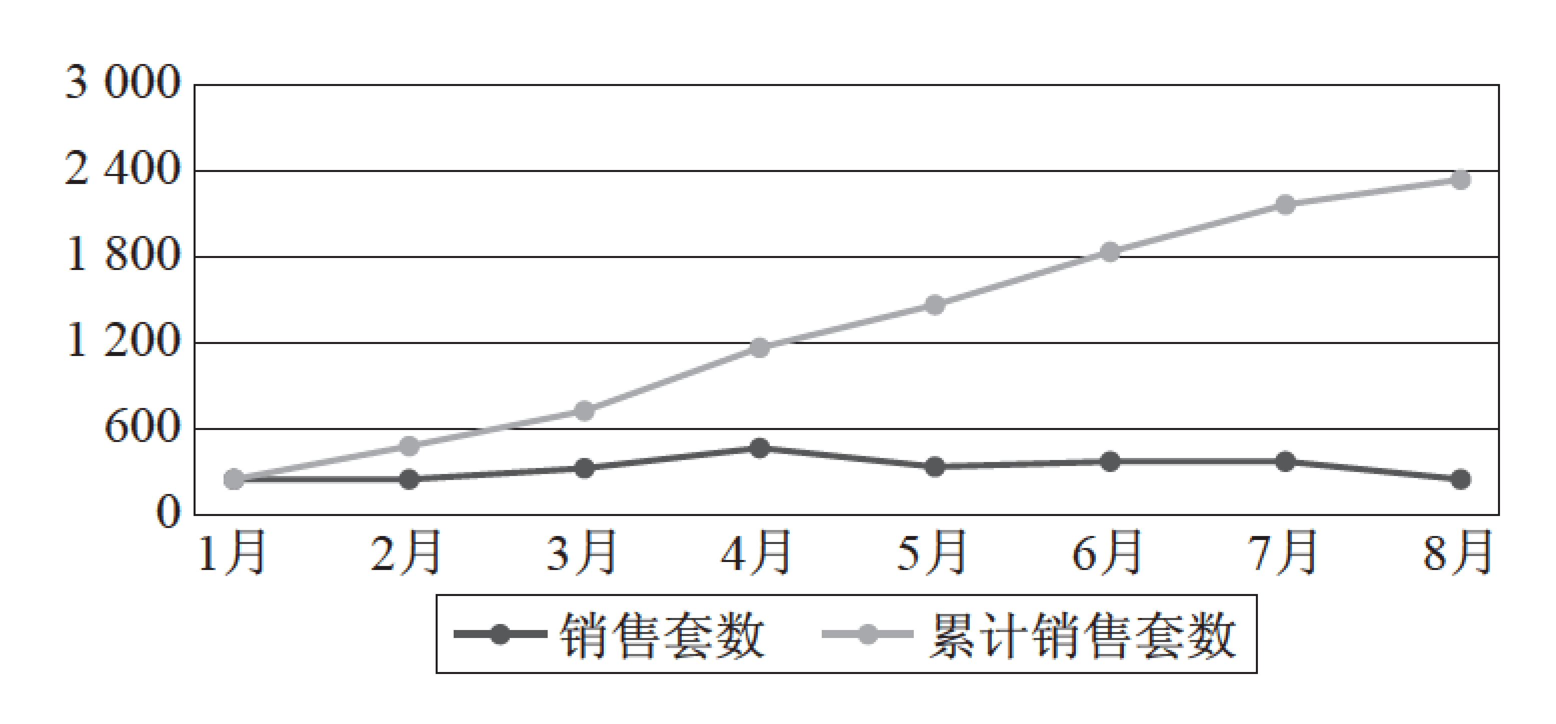


图 7-2 年度的时间走势（X 轴为月份）

通过钻取功能，可以查看一个月每天的销售套数的走势，如图 7-3 所示，钻取 8 月份，可以查看 8 月份每天的销售套数。

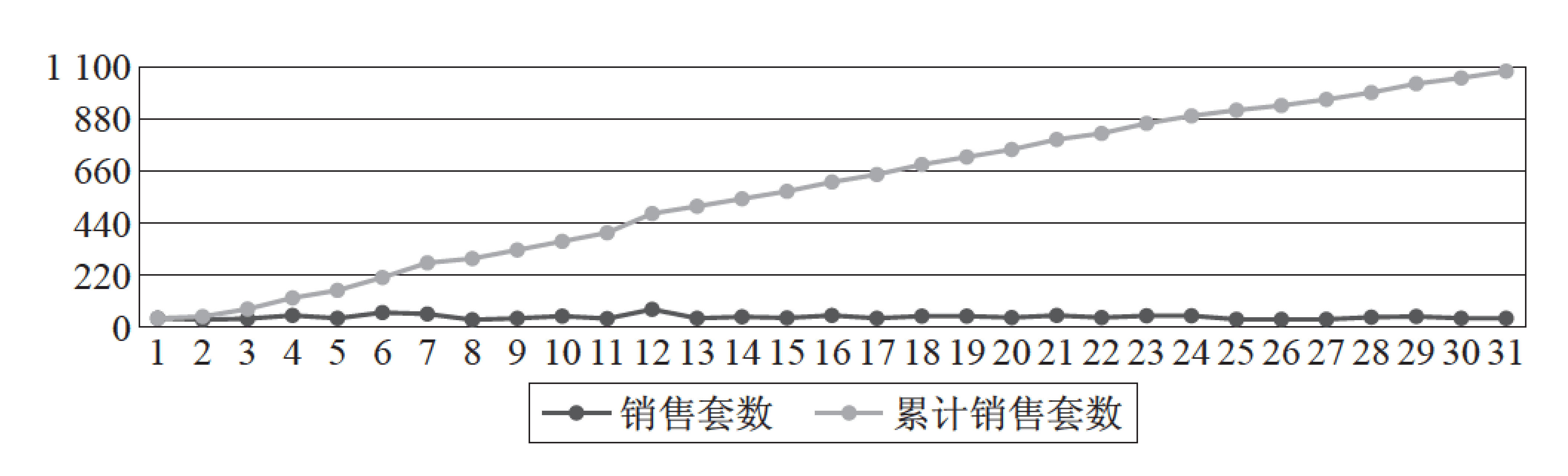


图 7-3 钻取到月份的时间走势（X 轴为日期）

项目维度标签页可以查询任意时间、任意面积分类、任意户型的各个板块各个项目的相关指标的对比情况，通过直方图和饼图的形式展示销售套数、销售面积的对比情况。通过饼图，可以查看各个板块关于销售套数（销售面积）的占比情况。点击钻取主城区，可以查看主城区下各个项目的销售套数（销售面积）的对比情况。通过饼图，可以查看各个板块关于销售套数（销售面积）的占比情况。

对于每个典型指标，都可以查看该指标的更加详细的分析数据。以项目维度的销售套数为例展示销售套数的本期值、同期值、同比增长率、环比增长率、销售套数最高的前 10 个项目及前 10 个项目的占比，前 10 个项目的时间走势、销售套数同比增长最快的前 10 个项目及同比增长最慢的前 10 个项目、销售套数环比增长最快的前 10 个项目及环比增长最慢的

前 10 个项目。

销售套数最高的前 10 个项目统计的是所有板块中销售套数最高的前 10 个项目以及这 10 个项目的上期销售套数、销售套数的环比增长率以及销售面积的情况，以多横条图的形式展示，使对比更加明显，如图 7-4 所示。

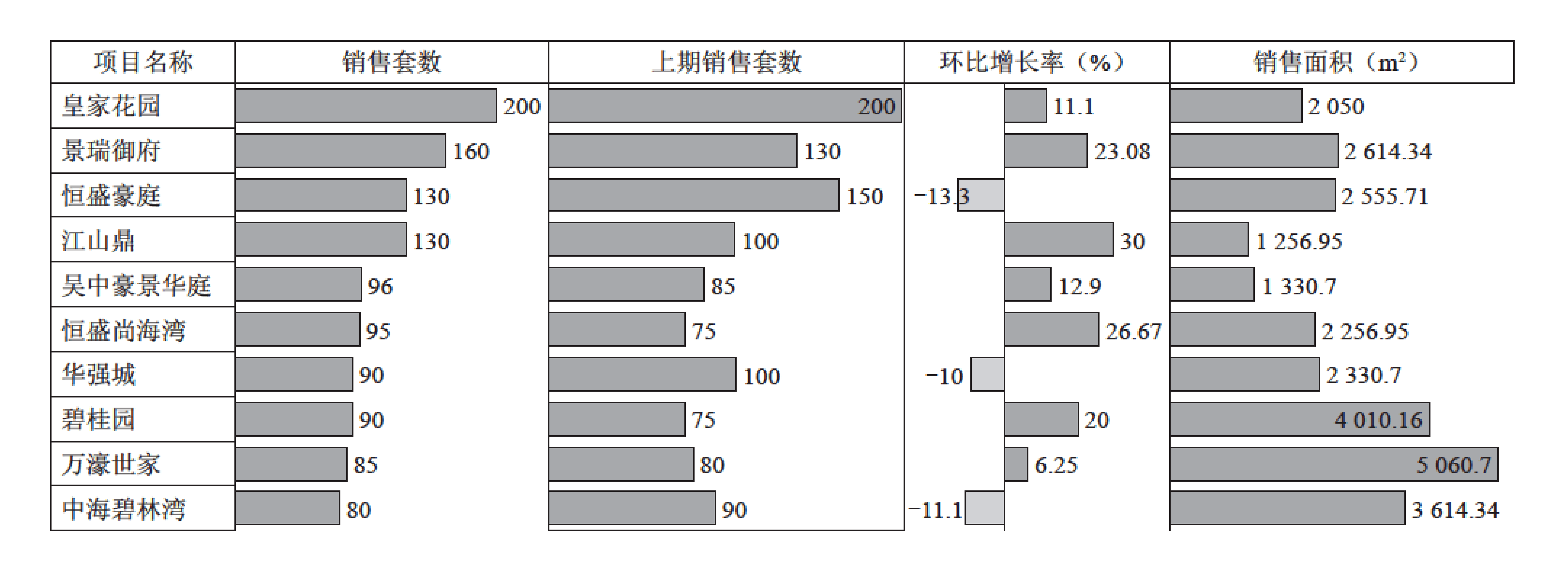


图 7-4　销售套数最高的前 10 个项目

面积分类维度标签页可以查询任意时间、任意板块、任意项目、任意户型的各个面积分类的相关指标的比较。通过直方图和饼图的形式分析销售套数、销售面积的比较，可以展示各个面积分类销售套数、销售面积的对比关系。通过饼图，查看各个面积分类的销售面积的占比情况。

户型维度标签页可以查看任意时间、任意板块、任意项目、任意面积分类的各个户型的相关指标的分析。通过直方图和饼图的形式分析销售套数、销售面积的比较。通过饼图，查看各个户型的销售套数的占比情况。

明细标签页展示了有关成交记录的明细数据，包括项目名称、楼盘地址、所在板块、面积、户型、物业类型、销售套数、销售面积的详细信息。

7.2　医院管理决策支持系统

医院决策支持系统分为两种类型：一是医院管理的决策支持，二是临床决策支持。

本方案主要用于医院管理，数据来源于医院的多个信息系统。医院管理决策支持系统以提供综合查询的监控中心作为进入其他分析主题的入口，起一个门户作用，显示最近一天的主要指标及随时间的变化，如果对某个指标感兴趣，可以点击链接进入相应的分析主题，从更多维度分析，并看到其他相关指标。

该方案从九个主题分析有关医院管理的数据，分别为医药收费数据分析、门诊数据分析、住院数据分析、用药数据分析、医疗项目收入数据分析、大型诊断检查数据分析、手术数据分析、体检数据分析和物资出入库数据分析。

7.2.1 监控中心

监控中心提供对最新数据和主要数据的监控和分析。

监控中心只与时间维度有关，与其他维度无关，查看任意日期的一些综合指标。监控中心从概要和时间两个维度标签页分析各个指标。

指标有基础指标和计算指标。监控中心的基础指标为门诊收费、住院收费、门诊量、门诊医疗收入、门诊药品收入、住院医疗收入、住院药品收入。

监控中心的计算指标有科室门诊收费 TOP10、科室住院收费 TOP10、医生门诊量 TOP10、医生门诊量 TOP10 份额、主要诊断门诊量 TOP10、主要诊断门诊量 TOP10 份额、科室门诊医疗收入 TOP10、科室门诊药品收入 TOP10、科室住院医疗收入 TOP10、科室住院药品收入 TOP10。其中科室门诊收费 TOP10 为门诊收费最高的前 10 个科室，科室住院收费 TOP10 为住院收费最高的前 10 个科室，医生门诊量 TOP10 为门诊量最多的前 10 位医生，医生门诊量 TOP10 份额为门诊量最多的前 10 位医生占所有医生总门诊量的占比情况，主要诊断门诊量 TOP10 为门诊量最高的前 10 种主要诊断，主要诊断门诊量 TOP10 份额为门诊量最高的前 10 种主要诊断占所有主要诊断的总门诊量的占比情况，科室门诊医疗收入 TOP10 为门诊医疗收入最高的前 10 个科室，科室门诊药品收入 TOP10 为门诊药品收

入最高的前 10 个科室，科室住院医疗收入 TOP10 为住院医疗收入最高的前 10 个科室，科室住院药品收入 TOP10 为住院药品收入最高的前 10 个科室。

概要标签页主要分析的是各个指标的合计数据以及相关指标的排名情况，用于实时监控医院有关收费以及门诊量的数据。概要页面分析的相关指标有：门诊收费、住院收费、门诊量、门诊医疗收入、门诊药品收入、住院医疗收入、住院药品收入，以上指标使用仪表盘表示。如图 7-5 所示，指针位置为门诊收费的本期值，红色和绿色交界的位置为门诊收费的上期值。

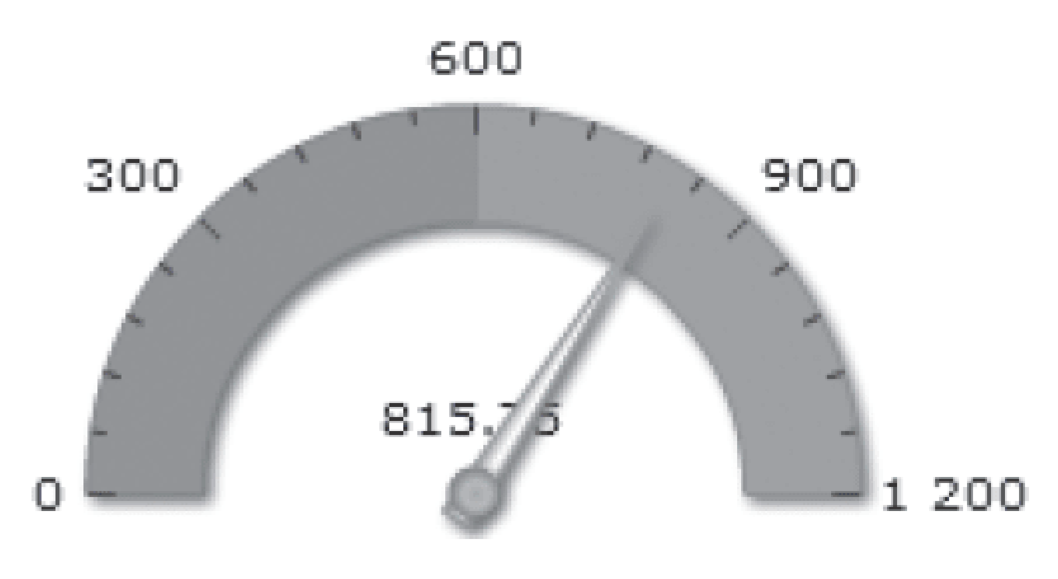


图 7-5　门诊收费的仪表盘

概要中还统计了相关指标的排名情况，有科室门诊收费 TOP10、科室住院收费 TOP10、医生门诊量 TOP10 以及份额情况、主要诊断门诊量 TOP10 以及份额、科室门诊医疗收入 TOP10、科室门诊药品收入 TOP10、科室住院医疗收入 TOP10、科室住院药品收入 TOP10。一般用横条图显示各种 TOP10 排名，通过饼图显示 TOP10 份额。

时间维度标签页可以查看各个年份，各个月份及每天的指标的走势情况。通过折线图分析门诊收费、住院收费、门诊量、门诊医疗收入、门诊药品收入、住院医疗收入、住院药品收入的走势情况。比如，将门诊收费和住院收费放在同一张折线图上，既可以分析门诊收费和住院收费的走势情况也体现了门诊收费和住院收费的对比关系。

7.2.2　医药收费数据分析

医药收费数据分析从时间、医师、医药、门诊住院、地区、性别、年龄以及病人属性 8 个维度分析关于收费的情况。

医药数据分析从时间、医师、医药、门诊住院、地区、性别、年龄段、病人属性及明细 9 个标签页面以及每个典型指标的二级页面分析收费指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

医药收费数据分析从时间、医师、医药、门诊住院、地区、性别、年龄段、病人属性 8 个维度分析各个指标。时间可以是就诊时间、入院时间、出院时间、手术时间，可以任意选择，时间维度的层次为年、月、日；医师维度的层次为科室、医生；门诊住院维度值为门诊或者住院；地区维度为病人所属地区为市区下面的各个区；性别维度值为男或者女；年龄段维度为各个年龄段，一般分为 0 ～ 6 岁、7 ～ 17 岁、18 ～ 40 岁、41 ～ 65 岁、66 岁以上；病人属性维度一般分为现金、儿童医保、居民医保、公务员医保、农保。

医药收费数据分析指标分为基础指标和计算指标，基础指标为收费，计算指标为同比增长率、环比增长率、指标值的排名前 10 名、指标值的排名后 10 名、指标值的份额。其中指标同比增长率为本期指标值与上一年同期的数据相比较的增长率，环比增长率为本期指标值对于本年上一期的数据相比较的增长率，指标值的排名前 10 名为一个维度下最后一个层次的某个指标的前 10 名，指标值的排名后 10 名为一个维度下最后一个层次的某个指标的后 10 名，指标份额是对于某个指标值占该指标合计值的占比。

时间维度可以查询任意科室、任意医生、任意医药、任意门诊住院、任意地区、男或者女、任意年龄段、任意病人属性的各个年份、各个月份及每天的关于收费的走势情况，通过折线图分析收费的走势情况。通过钻取功能，可以查看一个月每天的收费走势。

医师维度可以查询任意时间、任意医药、任意门诊住院、任意地区、男或者女、任意年龄段、任意病人属性的各个科室或者各个医生关于收费的对比情况。比如，通过直方图，将各个科室的收费放在一张图形上，更加明显地对比各个科室的收费情况。通过钻取功能，可以查询各个科室下每位医生的收费情况。

医药维度标签页可以查询任意时间、任意医师、任意门诊住院、任意地区、男或者女、任意年龄段、任意病人属性的有关医疗和药品的收费情况。

比如，通过直方图展示医疗和药品关于收费的对比情况，通过饼图，可以查看医疗和药品关于收费的占比情况。

门诊住院维度标签页可以查询任意时间、任意医师、任意医药、任意地区、男或者女、任意年龄段、任意病人属性的门诊、住院关于收费的对比情况，比如通过直方图的形式分析门诊和住院关于收费的对比情况，通过饼图的形式展示门诊和住院关于收费的占比情况。

地区维度标签页可以查询任意时间、任意医师、任意医药、任意门诊住院、男或者女、任意年龄段、任意病人属性的各个地区的关于收费的对比情况。比如，通过直方图的形式展示各个地区（以南通的各个地区为例）的收费情况，通过饼图可以查看各个地区关于收费的占比情况。

性别维度标签页可以查询任意时间、任意医师、任意医药、任意门诊住院、任意地区、任意年龄段、任意病人属性的男、女关于收费的对比情况。比如，通过直方图展示男、女关于收费的对比情况，通过饼图展示男、女关于收费的占比情况。

年龄段维度标签页可以查询任意时间、任意医师、任意医药、任意门诊住院、任意地区、男或者女、任意病人属性的各个年龄段的关于收费的对比情况。比如，通过直方图，展示各个年龄段的关于收费的对比情况，通过饼图可以查看各个年龄段关于收费的占比情况。

病人属性维度标签页可以查询任意时间、任意医师、任意医药、任意门诊住院、任意地区、男或者女、任意年龄段的各个病人属性的有关收费的对比情况。比如，通过直方图，展示了各个病人属性的关于收费的对比情况，通过饼图展示各个病人属性的关于收费的占比情况。

对于每个典型指标，还可以查看该指标更加详细的分析数据。以医师维度的收费指标为例，可以展示收费的本期值、同期值、同比增长率、上期值、环比增长率、收费最高的前10位医生以及前10位医生的收费的占比、前10位医生的收费的时间走势，收费同比增长最快的前10位医生及同比增长最慢的前10位医生、收费环比增长最快的前10位医生及环比增长最慢的前10位医生。

明细页面展示了有关医药收费的明细数据，包括日期、科室、医生、

医药类型、门诊住院、地区、年龄、病人属性、收费的详细信息。

7.2.3 门诊数据分析

门诊数据分析从时间、医师、主要诊断 3 个维度分析门诊人次、急诊人次、留观人次、医疗收入、医药收入、药占比等指标的走势及对比情况。

门诊数据分析从时间、医师、主要诊断 3 个维度分析指标，分为时间、医师、主要诊断、明细 4 个标签页面，以及每个典型指标的二级页面分析各个指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

时间维度层次为年、月、日，医师维度层次为科室、医生，主要诊断维度的层次为科室、主要诊断。

门诊数据分析基本指标有门诊人次、急诊人次、留观人次、医疗收入、药品收入。

计算指标有人均医疗收费、每人次药品收费、药占比、医均门诊人次、医均医药收入、指标的同比增长率、指标的环比增长率、指标排名前 10 名、指标排名后 10 名、指标份额。其中人均医疗收费等于总医疗收入除以总人数，每人次药品收费等于总药品收入除以总人次，药品占比等于药品收入与医疗收入和药品收入之和的比率，医均门诊人次为总的门诊人次除以医生人数，医均医药收入为药品收入除以医生总人数，其中指标同比增长率为本期指标值与上一年同期的数据相比较的增长率，环比增长率为本期指标值对于本年上一期的数据相比较的增长率，指标值的排名前 10 名为一个维度下最后一个层次的某个指标的前 10 名，指标值的排名后 10 名一个维度下最后一个层次的某个指标的后 10 名，指标份额是对于某个指标值占该指标合计值的占比。

时间维度标签页可以查看任意科室、任意医生、任意主要诊断的各个年份、各个月份以及每天的指标的走势情况。通过折线图或者时序堆积图分析门诊人次、急诊人次、留观人次、医疗收入、药品收入等指标的走势情况。通过时序堆积图，将医疗收入和药品收入放在一张图上面（见图 7-6），

既可以看出医疗收入和药品收入的走势情况也体现了二者的合计关系。

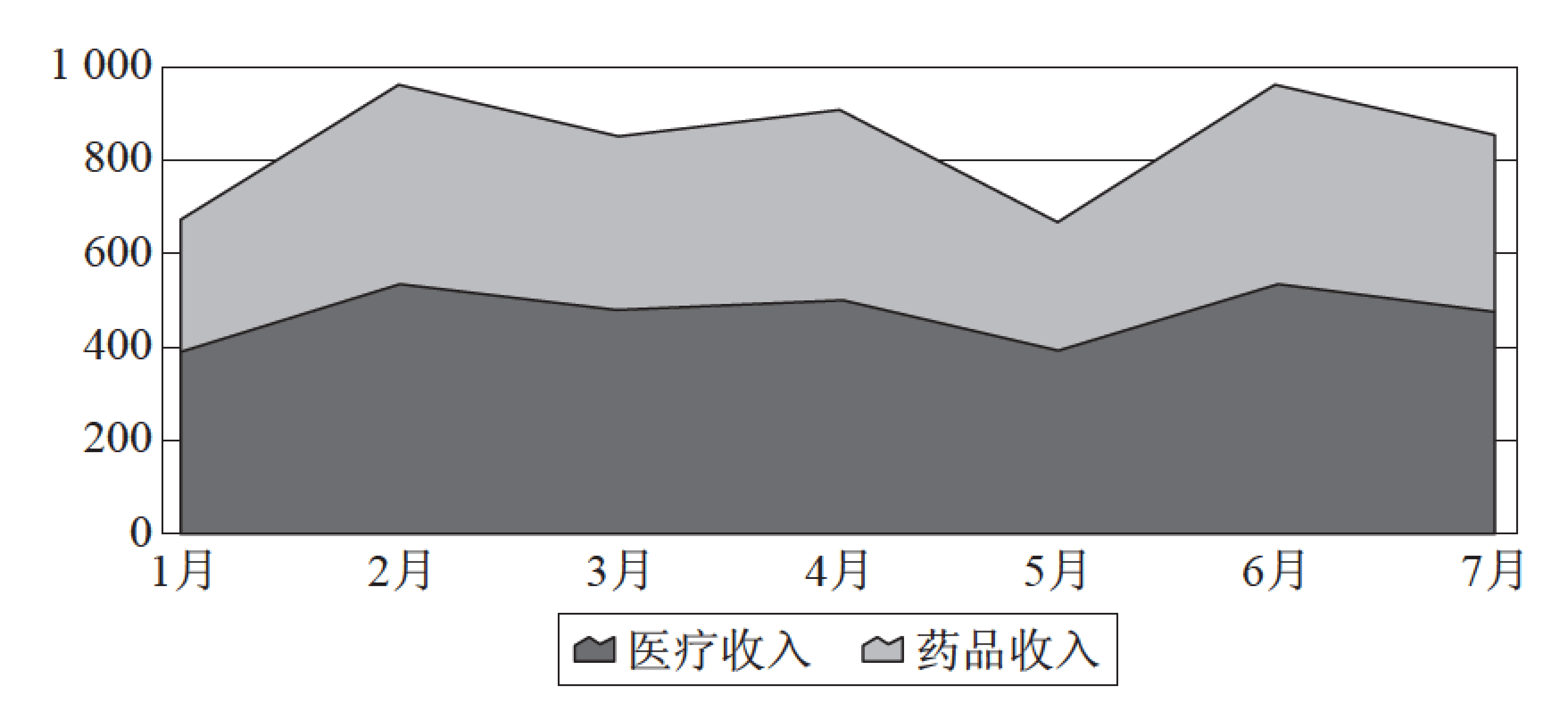


图 7-6　用堆叠图按月份显示的时间维度数据

通过钻取功能，可以查看每个月每天的数据走势情况。如图 7-7 所示，钻取 5 月份可以查看 5 月每天医疗收入和药品收入的走势及合计情况。

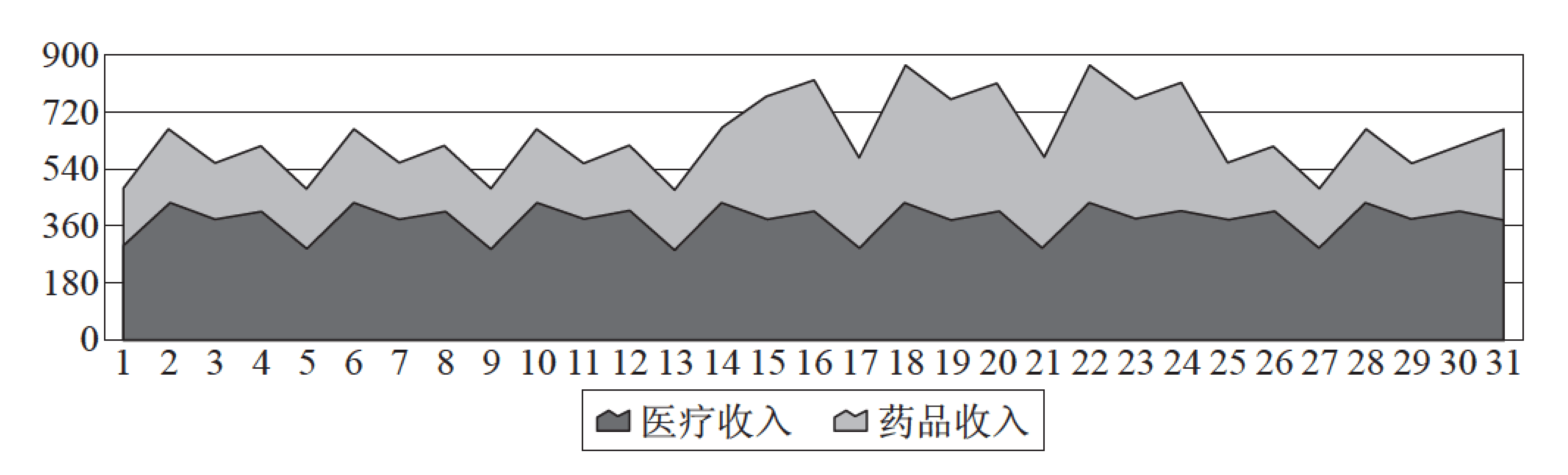


图 7-7　从月份钻取到日期后的时间维度数据

通过折线图体现指标的走势情况，将门诊人次、急诊人次、留观人次 3 个指标放在一张图形上，既可以查看这 3 个指标的走势情况也体现了这 3 个指标的对比情况。用钻取功能，可以查看每个月每天的数据走势，比如钻取 5 月，可以查看 5 月每天关于门诊人次、急诊人次、留观人次这 3 个指标的走势对比情况。

医师维度标签页可以查询任意时间、任意主要诊断的各个科室、每位医生的指标对比情况。通过直方图或者直方堆积图展示各个指标的对比情况。如图 7-8 所示，使用双 Y 直方堆积图，既体现了医疗收入和药品收入的合计情况，也体现了药占比的情况。

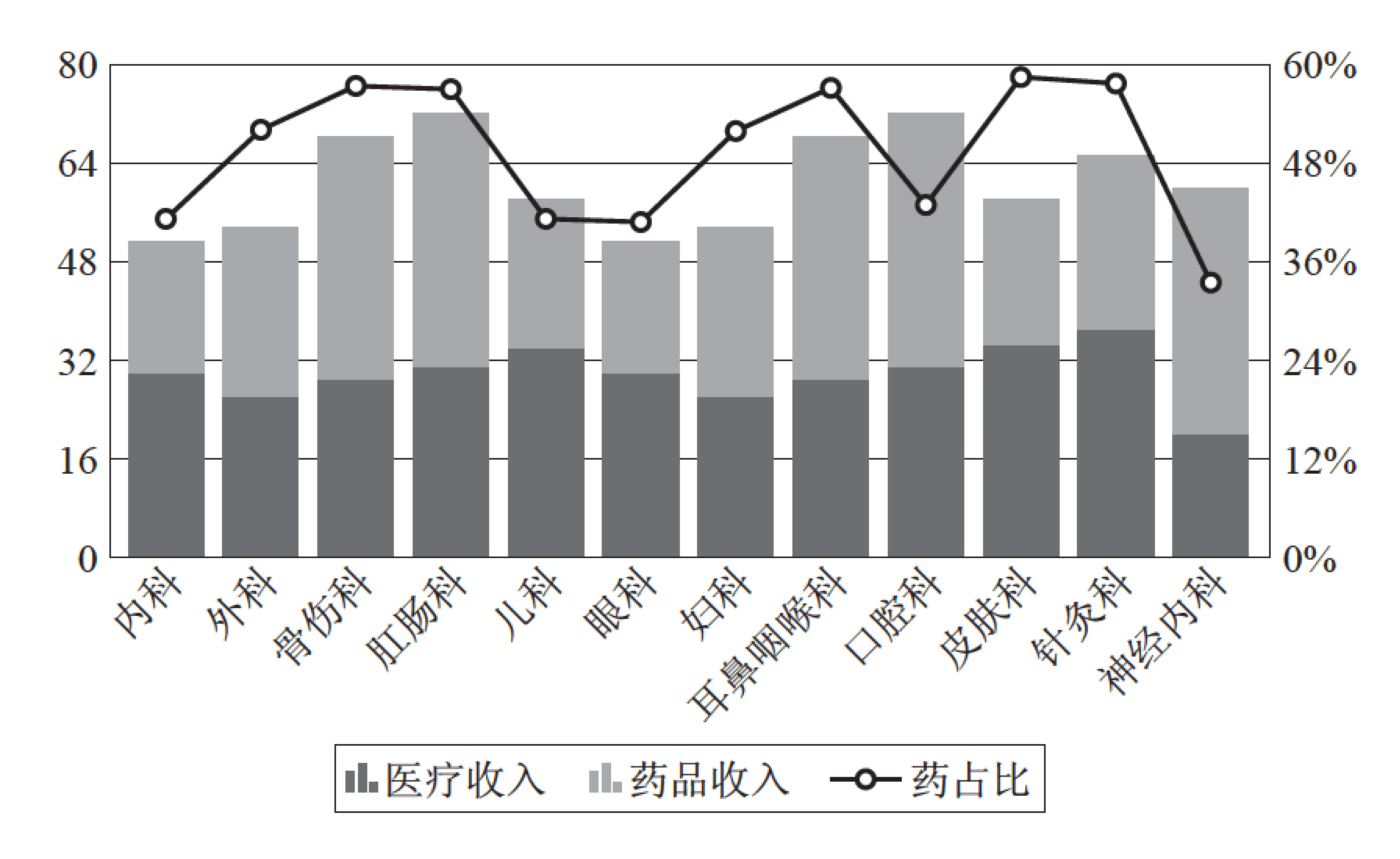


图 7-8　医师维度在科室级别的数据比较

通过钻取功能，可以查看每个科室每位医生各个指标的对比情况。如图 7-9 所示，钻取内科，可以查看内科每位医生的医疗收入、药品收入及药占比的对比情况。

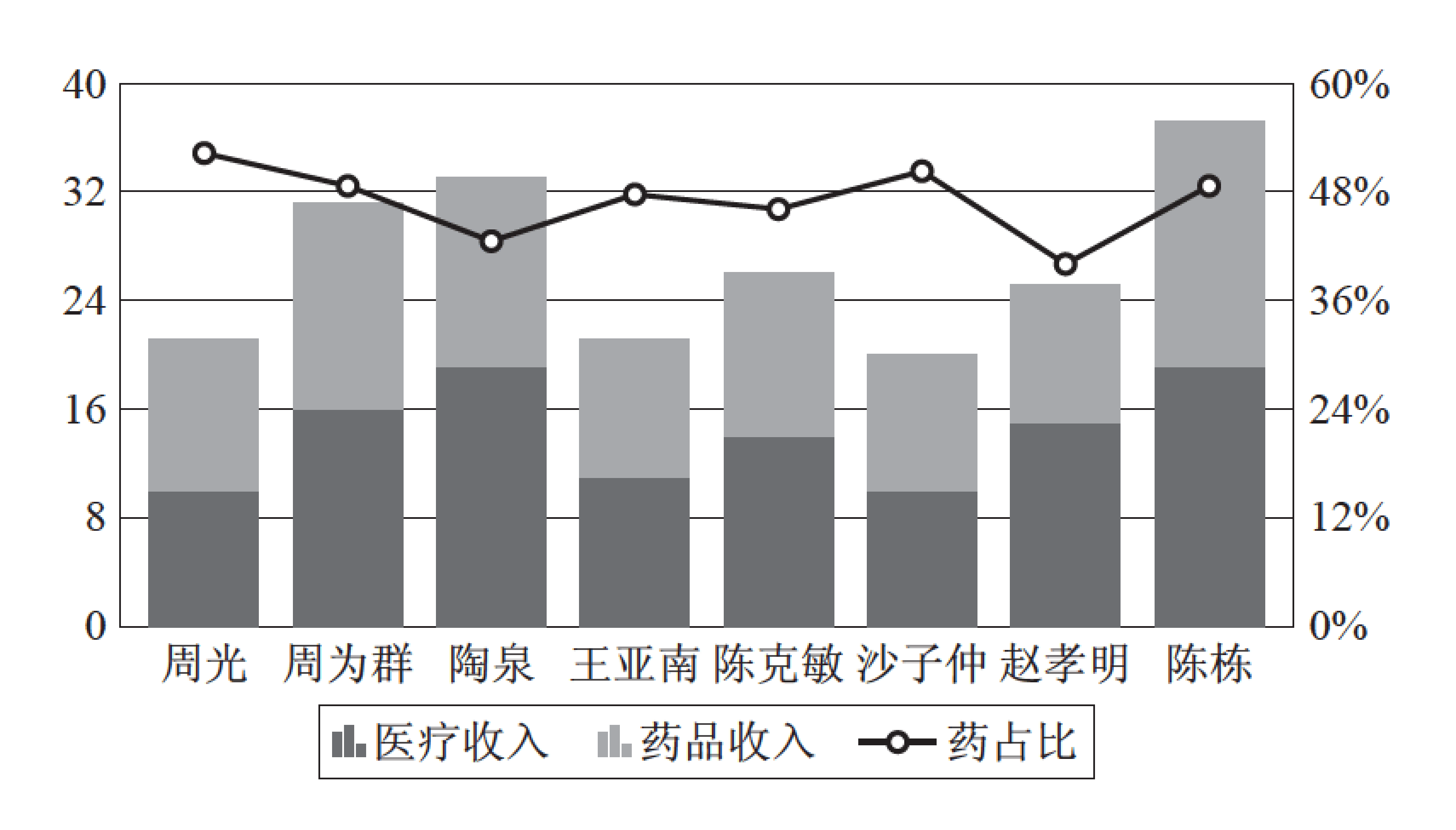


图 7-9　医师维度钻取到医生级别的数据比较

主要诊断维度标签页可以查询任意时间、任意医师的各个科室、各个主要诊断的指标对比情况。通过直方图或者直方堆积图展示各个指标的对比情况。如图 7-10 所示，通过直方图，将门诊人次、急诊人次、留观人次放在一张图形上，更加明显地体现了这 3 个指标在各个科室的对比关系。

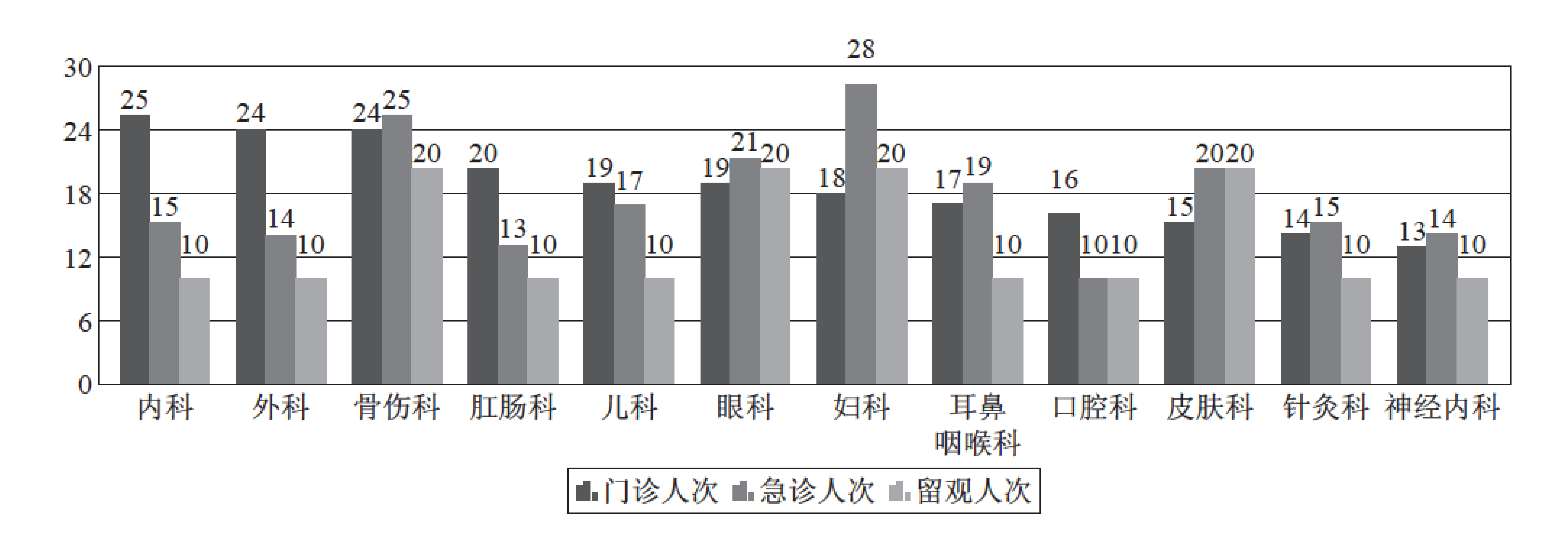


图 7-10　主要诊断维度在科室级别的数据比较

通过钻取功能，可以查看各个科室每个主要诊断的指标对比情况。如图 7-11 所示，钻取内科，可以查看内科每种主要诊断的门诊量。

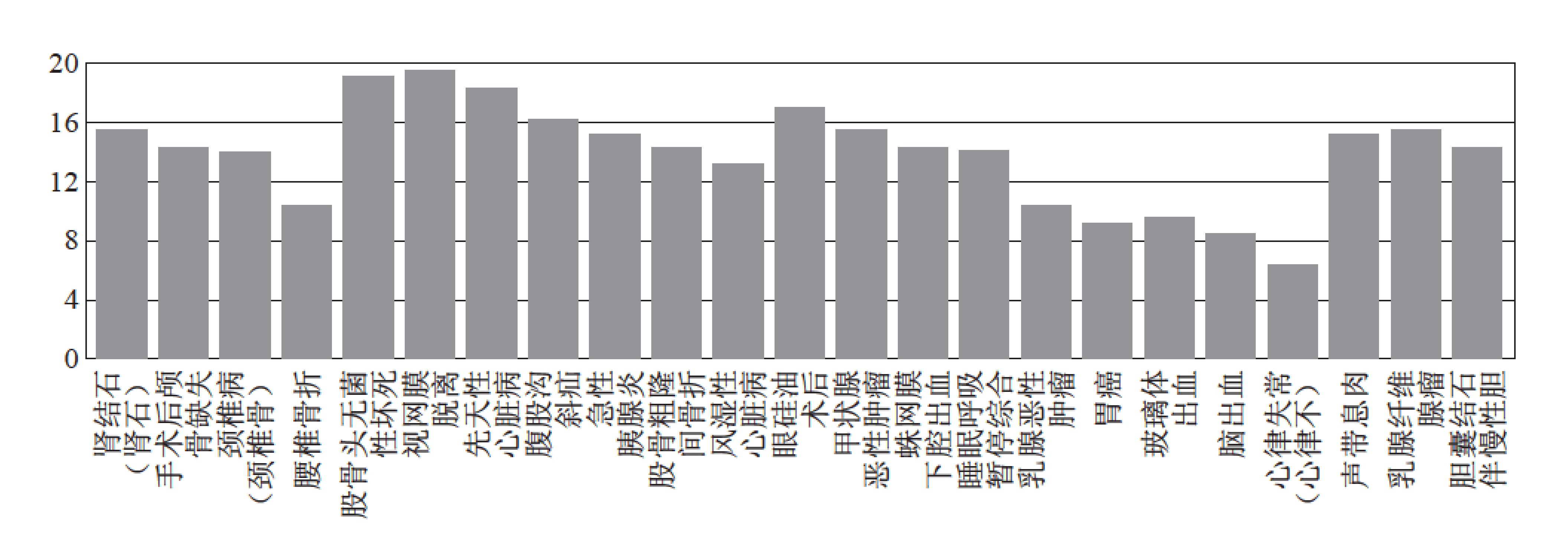


图 7-11　主要诊断维度钻取到诊断级别的数据比较

对于每个典型指标，都有一个对应的二级页面，可以查看该指标更加详细的分析数据。以主要诊断维度的门诊量为例展示门诊量的本期值、同期值、同比增长率、上期值、环比增长率，门诊量最高的 10 种病种及前 10 种主要诊断的占比，前 10 钟主要诊断的时间走势，门诊量同比增长最快的前 10 种主要诊断及同比增长最慢的前 10 种主要诊断，门诊量环比增长最快的前 10 种主要诊断及环比增长最慢的前 10 种主要诊断。

明细页面用类似报表的表格形式展示了有关门诊分析的明细数据，包括日期、科室、医生、主要诊断、门诊量、急诊人次、留观人次、医疗收入、药品收入。数据是被前面的各个维度联合过滤后的一个子集。

7.2.4 住院数据分析

住院数据分析从时间、医师、主要诊断、出院状态 4 个维度分析医疗收入、药品收入、药占比等相关指标。住院数据分析的分析方式同门诊数据分析类似。

住院数据分析从时间、医师、主要诊断、出院状态、明细 5 个标签页以及典型指标的二级页面分析医疗收入、药品收入、占用床位日等相关指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

指标分为基础指标和计算指标，其中基础指标包括医疗收入、药品收入、床位占用数、入院人数、出院人数、出院者占用床位日；计算指标包括药占比、出院者平均住院日、病床使用率、病床周转次数、每床日收费额、每床日收药品费、平均病床工作日。其中药占比为药品收入与医疗收入和药品收入之和的比率，出院者平均住院日为出院者占用总床日数除以出院人数，病床使用率为实际占用的总床日数与实际开放的总床日数之比，病床周转次数为出院人数除以平均开放病床数，每床日收费额为医疗收入和药品收入之和除以住院床日，每床日收药品费用为药品收入除以住院床日，平均病床工作日为实际占用总床日数除以平均开放病床数。

时间维度标签页可以查看任意科室、任意医生、任意主要诊断、任意出院状态的各个年份、各个月份以及每天的相关指标的走势情况。通过时序堆积图和折线图分析医疗收入、药品收入、药占比、占用床位日、平均住院日、入院人数、出院人数等指标的走势情况。

医师维度标签页可以查询任意时间、任意主要诊断、任意出院状态的各个科室、每位医生的医疗收入、药品收入、药占比、占用床位日、平均住院日、入院人数、出院人数等指标的对比情况。通过直方图或者双 Y 直方堆积图展示各个指标的对比情况。

主要诊断（科室 > 主要诊断）维度标签页可以查询任意时间、任意医师、任意出院状态的各个科室、每种主要诊断的医疗收入、药品收入、药占比、占用床位日、平均住院日、入院人数、出院人数等指标的对比情况。通过直方图或者双 Y 直方堆积图展示各个指标的对比情况。

出院状态维度标签页可以查询任意时间、任意医师、任意主要诊断的各个出院状态的关于医疗收入、药品收入、药占比、占用床位日、平均住院日、入院人数、出院人数等指标的对比情况。通过直方图或者双 Y 直方堆积图展示各个指标的对比情况。

每个典型指标，都有一个对应的二级页面，可以查看该指标更加详细的分析数据。主要分析指标的本期值、同期值、上期值、同比增长率、环比增长率以及相关指标的排名情况。

明细页面用类似报表的表格形式展示了相关住院数据分析的明细数据，包括日期、科室、医生、主要诊断、出院状态、医疗收入、药品收入、入院人数、出院人数、科室病床使用率、病床周转次数、平均每床日收费额、平均每床日收费药品费、平均病床工作日的详细信息。数据是被前面的各个维度联合过滤后的一个子集。

7.2.5 手术数据分析

手术数据分析从时间、医师、手术类型 3 个维度分析手术例数、手术费用、手术成功例数、手术治愈例数、手术成功率、手术治愈率、手术占用床位日等指标的走势及对比情况。

手术数据分析从时间、医师、手术类型、明细 4 个标签页面以及典型指标的二级页面分析各个指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

手术数据分析的维度包括时间、医师、手术类型三个维度。其中时间维度层次为年、月、日，医师维度层次为科室、医生，手术类型维度层次为科室、手术等级、手术名称。

手术数据分析的指标分为基础指标和计算指标，其中基础指标有手术例数、手术成功例数、手术治愈例数、手术医疗收入、手术药品收入、手术占用床位日；计算指标有药占比、手术成功率、手术治愈率、手术出院病人平均住院日、科室手术病床使用率、手术病床周转次数、每例手术药品费。其中药占比为手术药品收入与手术医疗收入和手术药品收入的比率，

手术成功率为手术成功例数与总手术例数的比率，手术治愈率为手术治愈例数与手术总例数的比率，手术出院病人平均住院日为手术出院者占用总床日数除以出院人数，科室手术病床使用率为手术实际占用总床日数除以手术实际开放总床日数，手术病床周转次数为手术人数除以平均开放病床数，每例手术药品费为手术药品费用除以手术例数。

时间维度标签页可以查询任意医师、任意手术类型的各个年份、各个月份以及每天的指标的走势情况。通过折线图展示各个指标的走势情况，如图 7-12 所示，体现了每天手术成功率和手术治愈率的走势情况。

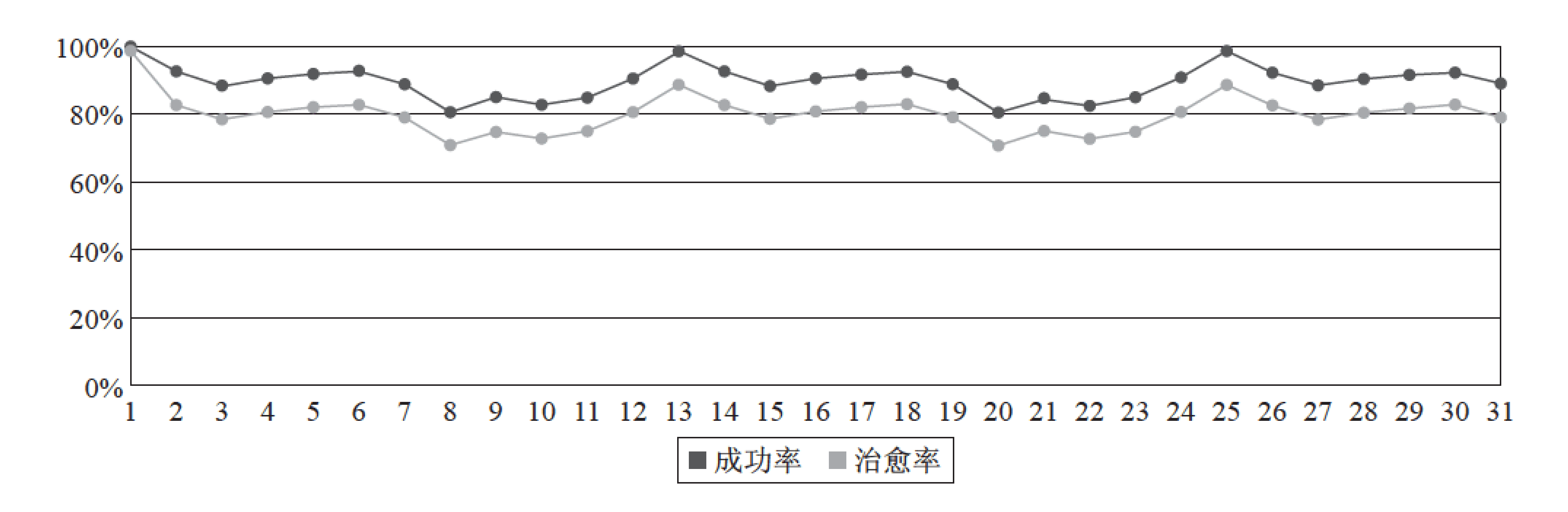


图 7-12　指标在时间维度的比较

医师维度标签页可以查询任意时间、任意手术类型的各个科室、各个医生的有关手术例数、手术医药收入、药占比、手术成功率、手术治愈率、手术占用床位日、手术出院病人平均住院日、科室手术病床使用率、手术病床周转次数、每例手术收药品费等指标的对比情况。如图 7-13 所示，通过直方图展示手术成功率和手术治愈率的对比情况。

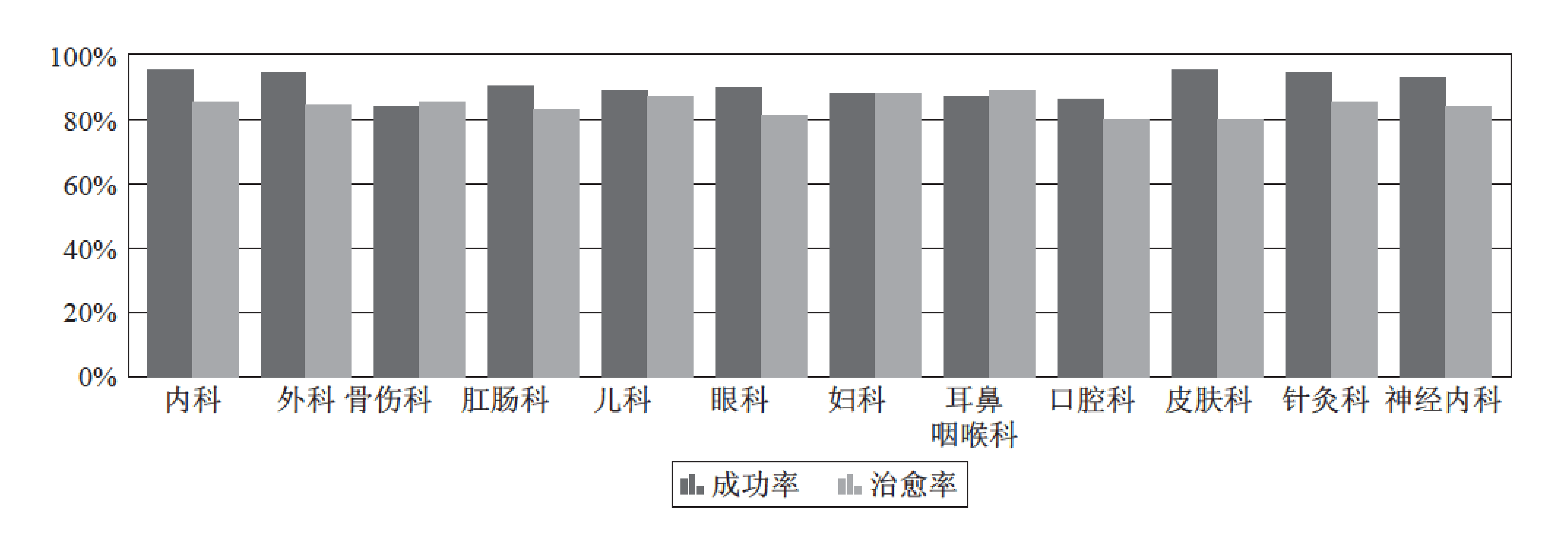


图 7-13　指标在科室维度的比较

手术类型维度标签页可以查询任意时间、任意科室、任意医生的各个手术类型的有关手术例数、手术医药收入、药占比、手术成功率、手术治愈率、手术占用床位日、手术出院病人平均住院日、科室手术病床使用率、手术病床周转次数、每例手术收药品费等指标的对比情况。

明细页面用类似报表的表格形式展示了有关手术数据分析的明细数据，包括日期、科室、医生、手术类型、手术等级、手术名称、手术医疗收入、手术药品收入的详细信息。

对于每个典型指标，都有一个对应的二级页面，可以查看该指标更加详细的分析数据。可以查看指标的本期值、同期值、上期值、同比增长率、环比增长率以及指标的排名情况。如图 7-14 所示，可以查看手术治愈率最高的 10 种手术以及这 10 种手术医疗收入、手术药品收入情况。

手术名称	手术治愈率	手术医疗收入	手术药品收入
阑尾切除术	100%	1 000	300
剖腹产术	100%	8 456	1 258
痔切除术	100%	2 014	1 058
小清创缝合术	100%	750	201
小腿骨折闭合复位术	99.45%	8 756	1 389
胆囊切除术	90%	5 764	1 560
疝修补术	85.48%	4 589.79	759.54
胰切除术	84.59%	4 897	1 025
韧带修补	80%	3 589	600
腱缝合术	80%	8 954	1 236.48

图 7-14　指标的排名比较

7.2.6　用药数据分析

用药数据分析主要从时间、医师、药品 3 个维度分析用药数量、用药金额、平均用药金额等指标的走势及对比情况。其中，时间维度的层次为年 > 月 > 日，医师维度的层次为科室 > 医生，药品维度的层次为药品种类 > 药品分类 > 药品名称。用药数据分析从时间、医师、药品、明细 4 个标签页面以及典型指标的二级页面分析各个指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

7.2.7 医疗项目收入数据分析

医疗项目收入数据分析从时间、医师、医疗收入类型3个维度分析数量、金额、平均金额的走势及对比情况。其中，时间维度的层次为年＞月＞日，医师维度的层次为科室＞医生，医疗收入类型的维度值有床位收入、挂号收入、护理收入、化验收入、检查收入、手术收入、诊察收入、治疗收入、其他收入。医疗项目收入数据分析从时间、医师、医疗收入类型、明细4个标签页以及典型指标的二级页面分析各个指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

7.2.8 大型诊断检查数据分析

大型诊断检查数据分析从时间（预约时间、就诊时间）、医师、检查类型3个维度分析检查次数、收费金额、平均收费等指标的走势及对比情况。其中时间可以为预约时间或者就诊时间，时间维度层次为年＞月＞日，医师维度的层次为科室＞医生，检查类型的维度值为CT、MRI、彩B、PET等。大型诊断检查数据分析从时间、医师、检查类型、明细4个标签页面以及典型指标的二级页面分析各个指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

7.2.9 体检数据分析

体检数据分析从时间、体检项目2个维度分析体检人数、异常人数、异常占比的走势及对比情况。其中时间维度的层次为年、月、日，体检项目维度值有血常规、内科检查、血脂、尿常规、妇科检查、一般检查、B超、防癌检查、外科检查、胸透、肾功能、肝功能。基础指标有体检人数、异常人数，计算指标异常占比为异常人数与体检人数之比。体检数据分析从时间、体检项目、明细3个标签页面以及典型指标的二级页面分析各个指标。对于多层次的维度标签页，可以通过钻取功能查看下一层维度页面的指标。

如体检项目维度标签页，通过直方双 Y 图体现体检人数、异常人数、异常占比的对比情况如图 7-15 所示。

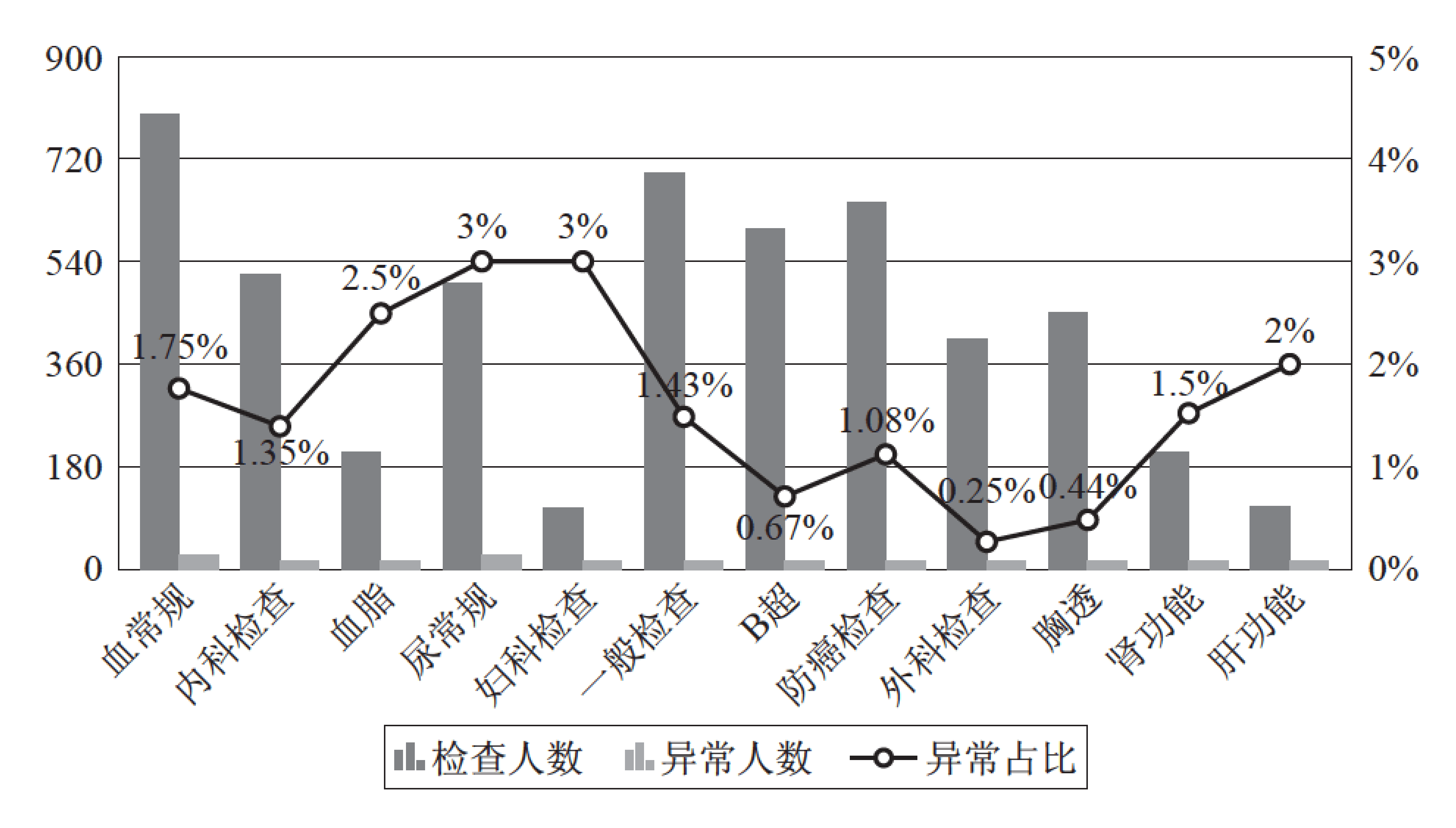


图 7-15　体检项目的一般比较

体检项目维度标签页，通过饼图展示各个体检项目体检人数的占比情况，如图 7-16 所示。

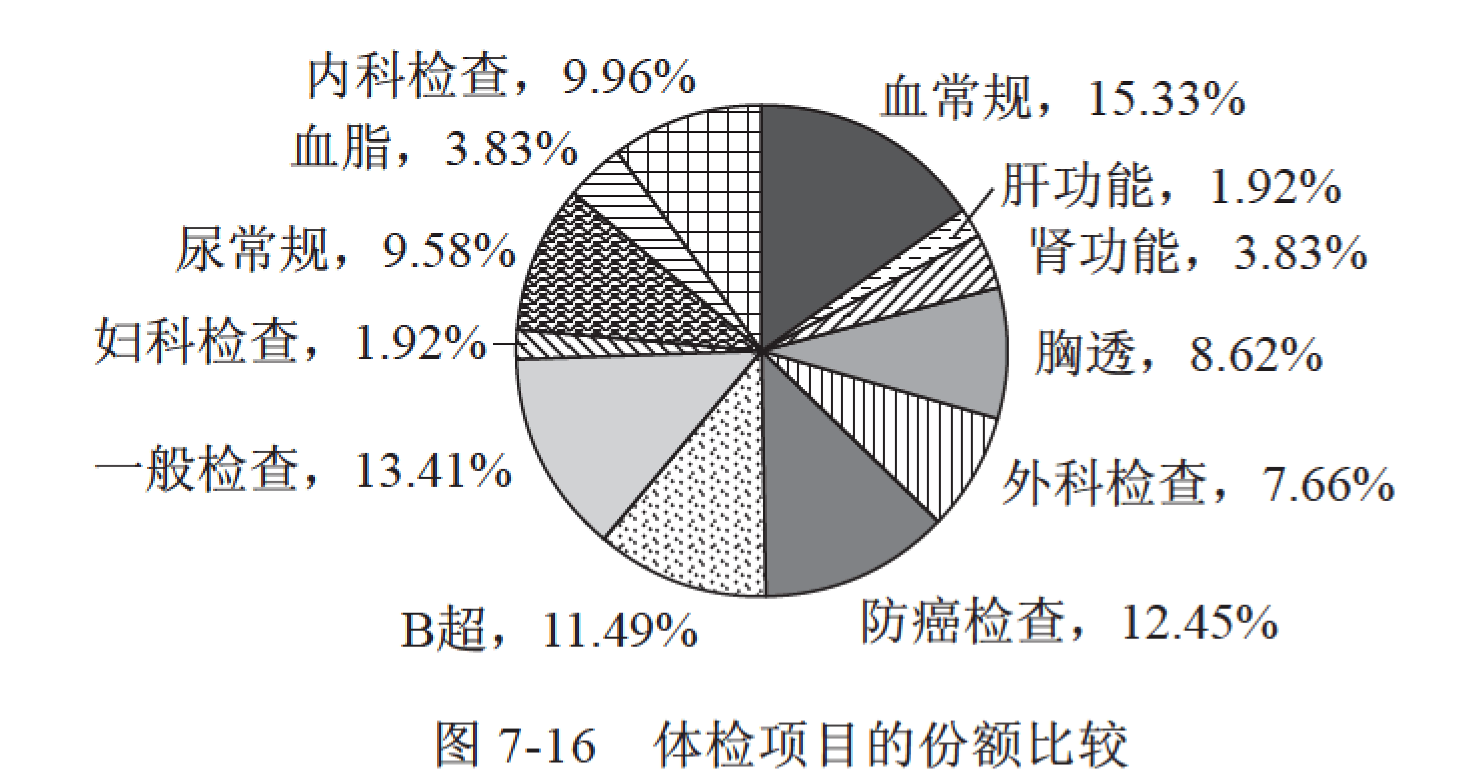


图 7-16　体检项目的份额比较

7.2.10　物资出入库数据分析

物资出入库数据分析从时间、科室、物资类别 3 个维度分析入库数量、出库数量、现存量、入库金额、出库金额的走势及对比情况，其中物资类

别的维度值有办公用品、妇科用药、固定资产、泌尿系统用药、生活用品、食品饮料、试剂用品、特殊医用材料、一次性卫生材料、医用低值易耗品。物资出入库数据分析从时间、科室、物资类别、明细 4 个标签页以及典型指标的二级页面分析各个指标。如在时间维度标签页，通过折线图将入库数量和出库数量放在同一张图上，体现了入库数量和出库数量的走势及对比情况，如图 7-17 所示。

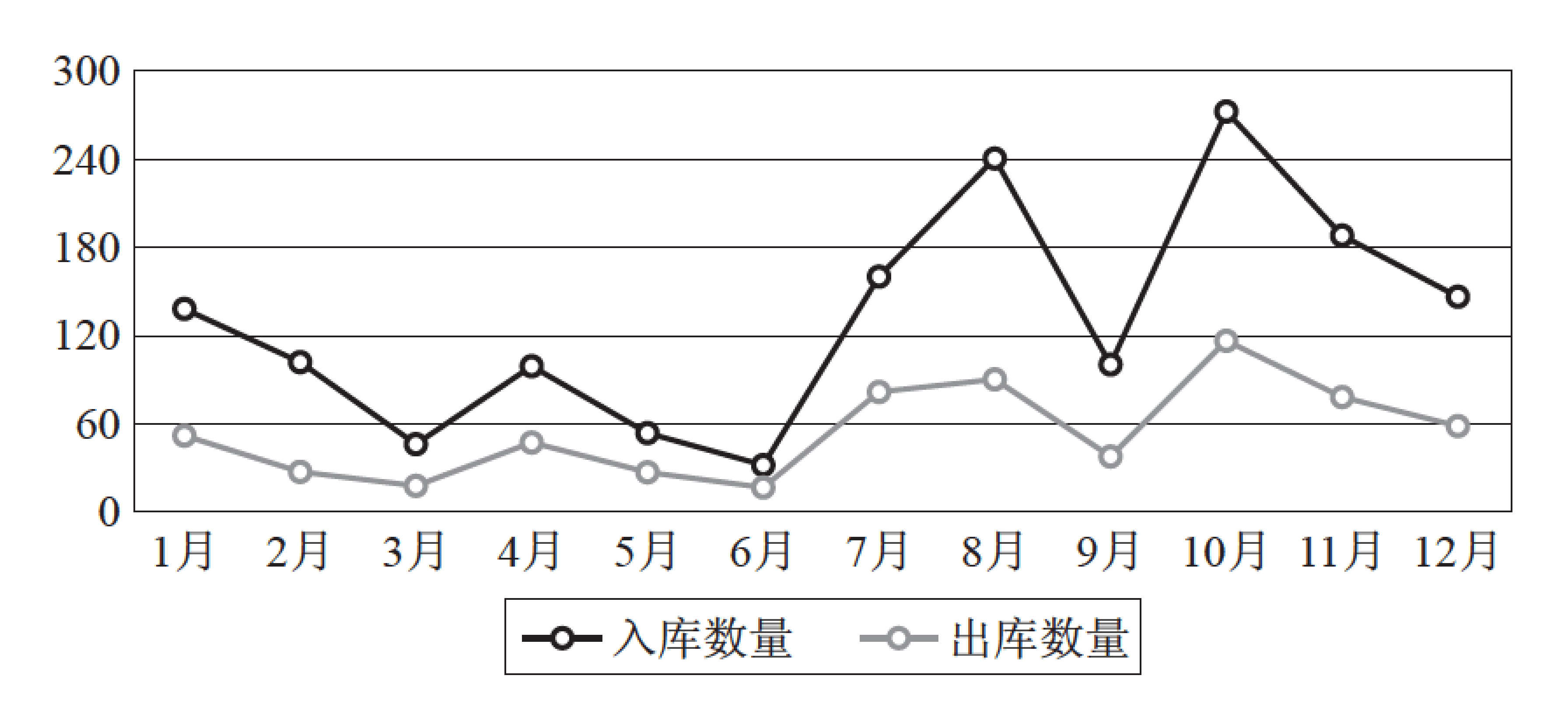


图 7-17　入出库的时间比较

在物资类别维度标签页，如图 7-18 所示，通过饼图展示各个物资类别的现存量的占比情况。

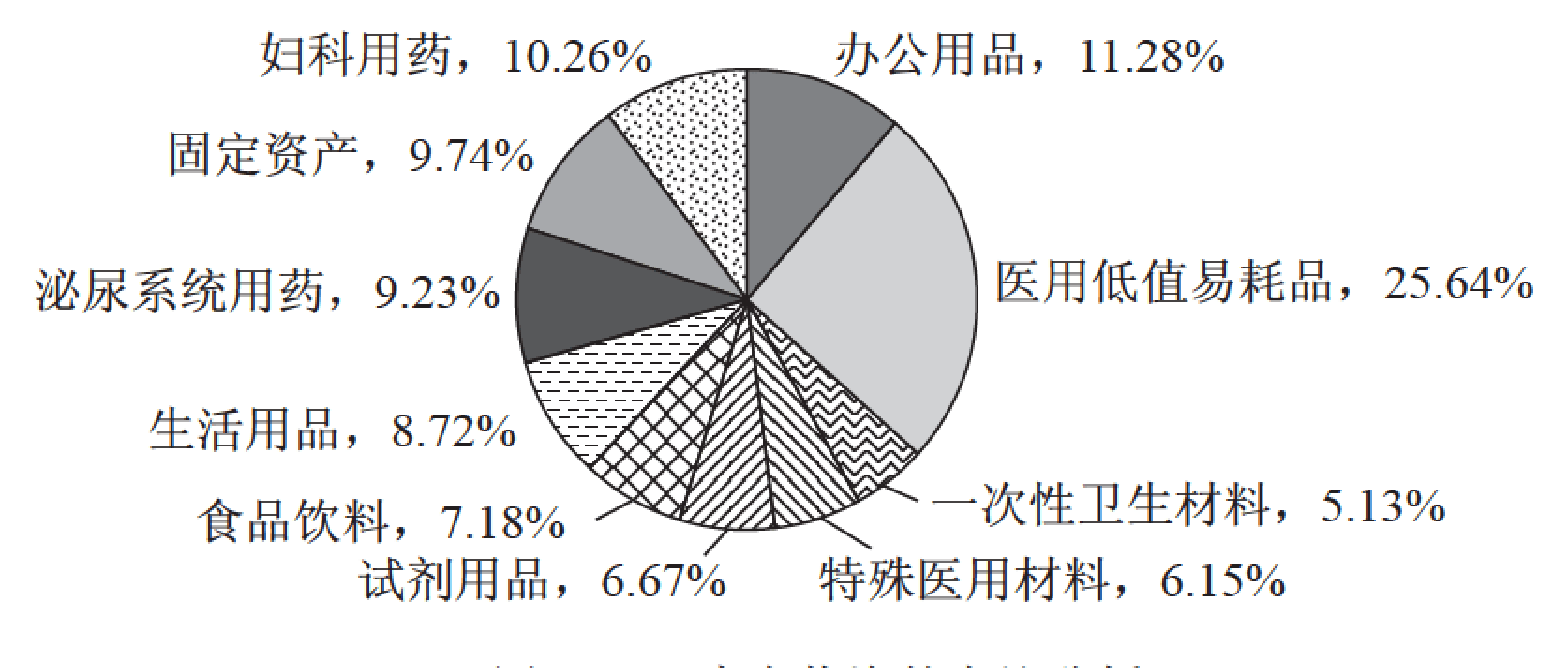


图 7-18　库存物资的占比分析

该方案从 6 个方面分析关于财政的相关数据，分别为监控中心、收入数据分析、支出数据分析、收支执行数据分析、预算执行用款分析、政府采购数据分析。

7.3 政府财政数据分析

经过多年信息化建设和推广，各地财政部门已经积累了大量丰富的财政业务数据资源，应该将这些资源进行整合，形成可用作分析决策的自助分析系统。

7.3.1 监控中心

监控中心提供对最新数据和主要数据的监控和分析，用于市政府和财政部门领导日常监控财政收支情况，并作为深入精细数据分析的门户。

监控中心的数据只与时间维度有关，与其他维度无关，可以查看任意日期的一些综合指标。

监控中心的指标有基础指标和计算指标，基础指标有预算收入、收入总金额、总支出预算、支出总金额、采购预算、采购金额。计算指标有单位收入总金额 TOP10 排名及份额、单位支出总金额 TOP10 排名及份额，其中单位收入总金额 TOP10 排名为总收入金额最高的前 10 个单位，单位收入总金额 TOP10 份额为总收入金额最高的前 10 个单位的收入金额占所有单位总收入的占比，单位支出总金额 TOP10 排名为支出总金额最高的前 10 个单位，单位支出总金额 TOP10 份额为支出总金额最高的前 10 个单位的支出金额占所有单位总支出的占比。

监控中心从概要和时间两个维度标签页分析各个指标。

概要标签页主要分析的是各个指标的合计数据以及相关指标的排名情况，用于平时监控有关财政收入、财政支出的相关数据。概要页面分析的相关指标有预算收入、收入总金额、总支出预算、支出总金额、采购预算、采购金额。通过仪表盘表示各个指标的合计数据，如通过仪表盘表示预算收入，仪表盘的指针位置为预算收入，仪表盘的颜色交界处为上年的决算值，将两个指标放在一张图形上，使对比更加明显。通过横条图分析相关

指标的排名情况，通过饼图展示相关指标排名的份额情况。

时间维度标签页可以查看各个年份及各个月份指标的走势情况。通过折线图分析预算收入、收入总金额、总支出预算、支出总金额、采购预算、采购金额的走势情况。可以将收入总金额和支出总金额放在一张图形上，既可以查看这两个指标的走势，也可以分析二者的对比关系。

7.3.2 收入数据分析

收入数据分析从时间、单位、科目三个维度分析预算收入、收入总额等指标的走势及对比情况。

收入数据分析从时间、单位、科目三个维度分析各个指标。时间维度的层次为年、月，单位维度层次结构为城市、局、下属单位，局如公安局、教育局、环保局、城乡建设局、房管局、交通运输局、农委、水利局等，科目为各个收入科目，如营业税、企业所得税、个人所得税、城市建设税、房产税、印花税、土地增值税、教育附加专项收入、地方财政税费附加收入等科目。

收入数据分析的指标有基础指标和计算指标，基础指标有年初预算收入、调整后预算、收入总额，计算指标有累计收入总额、收入占预算收入比、同比增长率、环比增长率、单位收入 TOP10 排名及份额、单位收入 BOTTOM10 排名等。其中累计收入总额为年初至所选日期的累计收入总额，收入占预算收入比为收入总额与调整后预算的比，同比增长率为本期指标值与上一年同期的数据相比较的增长率，环比增长率为本期指标值对于本年上一期的数据相比较的增长率，单位收入 TOP10 排名为收入最高的前 10 个单位，单位收入 TOP10 份额为收入最高的前 10 个单位的收入占所有单位总收入的占比，单位收入 BOTTOM10 排名为单位收入最低的 10 个单位。

收入数据分析有时间、单位、科目、明细 4 个标签页，以及每个基本指标的二级页面。对于多层次的维度标签页，可以通过向下钻取功能查看层次结构下一级的数据。

时间（年、月）维度标签页可以查看任意单位、任意科目的各个年份、各个月份以及每天的指标走势。通过折线图分析年初预算收入、调整后预算比较、收入总额、收入占预算收入等指标的走势情况。

单位维度标签页可以查看任意时间、任意科目的各个单位指标的对比情况。通过直方图分析年初预算收入、调整后预算比较、收入总额、收入占预算收入等指标的对比情况。

科目维度标签页可以查看任意时间、任意单位的各个科目指标的对比情况。通过直方图分析年初预算收入、调整后预算比较、收入总额、收入占预算收入等指标的对比情况。

对于每个基本指标，都有一个对应的二级页面，可以查看该指标的更加详细的分析数据。主要分析的是指标的本期值、同期值、上期值、同比增长率、环比增长率以及相关指标的排名情况。

明细页面以表格形式展示了相关收入数据的每笔明细，包括时间、单位名称、科目、年初预算收入、调整后预算收入、收入总金额。

7.3.3 支出数据分析

支出数据分析从时间、单位—项目、经济科目、来源科目 4 个维度分析年初预算支出、调整后预算支出、总支出预算、支出总额、预算执行进度等指标的走势及对比情况。

支出数据分析有时间、单位—项目、经济科目、来源科目 4 个维度，时间维度的层次为年、月，单位—项目的维度层次为单位、项目，经济科目的维度层次为经济科目、二级科目。

收入数据分析的指标有基础指标和计算指标，基础指标有年初预算支出、调整后预算支出、总支出预算、支出总金额、项目个数，计算指标有预算执行进度，其中预算执行进度计算为支出总金额与总支出预算的比。

支出数据分析有时间、单位—项目、经济科目、来源科目、明细 5 个标签页及基本指标的二级页面。对于多层次的维度，可以通过钻取功能查看下一级的数据。可以通过钻取经济科目查看该经济科目的来源科目，看

到某个经济科目的来源，也可以通过钻取来源科目查看各个经济科目，查看某个来源科目用在哪些经济科目上。

时间维度标签页，通过折线图展示各个指标的走势情况，将总支出预算和支出总金额放在一张图形上，展示了这两个指标的走势情况也体现了二者的对比关系。

单位—项目（单位、项目）维度标签页，通过直方图以及直方双 Y 图展示各个单位的相关指标的对比情况。通过直方双 Y 图，将总支出预算、支出总金额、预算执行进度放在同一张图形上，更加清晰地分析各个单位支出预算的执行情况。还可以通过钻取功能查询任意单位的各个项目的支出情况。

经济科目（科目、二级科目）维度标签页，通过直方图展示各个经济科目的关于总支出预算、支出总金额、预算执行进度的对比情况。通过钻取功能，查询经济科目下的各个二级科目的预算支出情况，同时还可以查询各个经济科目的来源科目。

来源科目维度标签页，通过直方图双 Y 图展示各个来源科目的关于总支出预算、支出总金额、预算执行进度的对比情况。还可以通过钻取功能，查询各来源科目用在了哪些经济科目上。

明细标签页用表格展示了支出数据的每笔明细，包括时间、单位名称、项目名称、经济科目、来源科目、年初预算、调整后预算、总支出预算、支出总金额。

7.3.4　收支执行数据分析

收支执行数据分析从时间、单位两个维度分析期初结余、收入总金额、支出总金额、期末结余指标的走势及对比情况。时间维度的层次为年、月。收支执行数据分析从时间、单位、明细 3 个标签页及基本指标的二级页面分析各个指标，其中明细页面用表格展示了收支数据的每笔明细，包括时间、单位名称、期初结余、收入总金额、支出总金额、期末结余。

7.3.5 预算执行用款数据分析

预算执行用款数据分析从时间、单位—项目两个维度分析用款支付节点发生的天数和次数。时间维度的层次为年、月，单位—项目的维度层次为单位、项目，指标有已批复的次数和天数、已生成支付令的次数和天数、汇总结算单形成的次数和天数、数据已转发银行的次数和天数、银行已回单的次数和天数、国库处已回单确认的次数和天数。预算执行用款数据分析从时间、单位—项目、明细 3 个标签页面分析各个指标，其中时间页面使用折线图分析任意单位、任意项目各个年份或者各个月份的用款支付节点的次数和天数的走势情况，单位—项目页面使用直方图分析任意时间各个单位、各个项目的用款支付节点的次数和天数的对比情况，明细页面展示的是预算执行用款数据的明细，包括时间、单位名称、项目名称、已批复天数、已生成支付令天数、汇总结算单形成天数、数据已转发银行天数、银行已回单天数、国库处已回单确认天数。

7.3.6 政府采购数据分析

政府采购数据分析从时间、单位、采购目录、供应商 4 个维度分析有关采购次数、采购预算总额、采购金额相关指标的走势及对比情况。

政府采购数据分析的维度有时间、单位、采购目录、供应商 4 个维度，时间维度层次为年、月，单位维度层次为单位、下属单位，采购维度层次为采购类别、货品分类、……、货品名称。对于多层次的维度可以通过钻取功能查看下一层次维度的指标。

政府采购数据分析的指标有基础指标和计算指标，基础指标为采购次数、采购预算总额（该指标与供应商无关）、采购金额，计算指标为采购次数占比、采购金额占比，累计采购预算总额、累计采购金额、采购次数占比为某个单位的采购次数与所有单位采购总次数的比值，采购金额占比为某个单位的采购金额与所有单位采购总金额的比值，累计采购预算总额为年初至今到所选日期的采购预算总额的合计值，累计采购金额为年初至

今到所选日期的采购金额的合计值。

政府采购数据分析从时间、单位、采购目录、供应商、明细 5 个标签页及每个基本指标的二级页面分析各个指标。

时间（年、月）维度标签页可以查询任意单位、任意采购目录、任意供应商的各个年份各个月份关于采购次数、采购预算总额、采购金额等指标的走势情况。通过折线图将采购预算总额和采购金额放在一张图形上既展示了这两个指标的走势情况，也体现二者的对比关系。

单位（单位、下属单位）维度标签页可以查询任意时间、任意采购目录、任意供应商的各个单位的关于采购次数、采购预算总额、采购金额的对比情况。通过直方图分析各个单位关于采购次数、采购预算金额、采购金额的对比关系，通过饼图体现各个单位关于采购次数、采购金额的占比情况。

采购目录（采购类别、货品分类、……、货品名称）维度标签页可以查询任意时间、任意单位、任意供应商的各个采购目录的关于采购次数、采购预算总额、采购金额的对比情况，通过直方图展示指标的对比情况。

供应商维度标签页可以查询任意时间、任意单位、任意采购目录的各个供应商的采购次数、采购金额的对比情况，通过直方图展示指标的对比情况。

对于每个基本指标，都有一个对应的二级页面，可以查看该指标更加详细的分析数据。主要分析的是指标的本期值、同期值、上期值、同比增长率、环比增长率以及相关指标的排名情况。如各个单位采购金额的同比增长率、环比增长率，采购金额最高的 10 个单位，采购金额最高的 10 种货品。

采购数据分析中的明细页面有两个，分别为政府采购明细和采购供应商明细，政府采购明细用表格显示时间、单位名称、采购商品、政府采购预算金额、采购金额的每笔数据；采购供应商明细用表格显示时间、单位名称、采购商品、采购金额的每笔数据。数据是被前面的各个维度联合过滤后的一个子集。

致　　谢

本书在写作中，得到不少人的帮助，在此一并致谢。

上海信息化发展研究协会徐龙章秘书长认真浏览了书稿，提出几个很好的修改意见，并提供一些参考资料。上海市软件行业协会杨根兴常务副会长浏览书稿并提出宝贵意见。南京航空航天大学丁秋林教授非常关心本书的出版。

王美玲参与文字整理，林彬彬整理了供应链的功能，郭晓杰整理了财务比率和案例。



参 考 文 献

[1] 刘红 . 大数据：第二次数据革命 [N]. 中国社会科学报，2014-1-21.

[2] 高中路 . 战后日本的解散财阀 . 外国问题研究，1986（2）.

[3] 大河网 - 大河报，明代藩王庞大的寄生集团，http：//news. ifeng. com/a/20150324/43403289_0. shtml，2015-3-24.

[4] 黄琪轩，李晨阳 . 从大国的市场开拓历史看中国“一带一路”战略 . http：//www. dfdaily. com/html/8762/2016/7/5/1360367. shtml，2016-07-05.

[5] Timothy H. Vines 等 . The Availability of Research Data Declines Rapidly with Article Age. http：//www. cell. com/current-biology/fulltext/S0960-9822(13)01400-0，2013-12-19.

[6] 荫蒙 . 数据仓库 [M]. 王志海等，译 . 北京：机械工业出版社，2006.

[7] 唐源，吴丹 . 国外医学科学数据共享政策调查及对我国的启示 [J]. 图书情报工作，2015（9）.

[8] 曹昌，李永华 . 医药流通业：阿里控制了零售数据不反对就没活路 [OL].

[9] 涂子沛，大数据 [M]. 桂林：广西师范大学出版社，2012.

[10] 人称 T 客 . Gartner 2016 年商业智能和分析平台魔力象限报告 [OL]. http：//iteyes. baijia. baidu. com/article/471859，2016-5-27.

[11] 大卫 • 麦克坎德雷斯 . 数据视觉化之美 [OL]. http：//open. 163. com/movie/2011/12/8/I/M8H1NPQM9_M8H1TQE8I. html，2011.

[12] Wiley. 统计学：可视化数据 . http：//v. 163. com/special/statisticsintroduction/.

[13] 金奇 . 金融信息不可靠困扰中国经济 [N/OL]. http：//www. ftchinese. com/story/001068493，2016-7-18.
[14] [美] 艾伦 • 艾伯斯坦 . 秋风译 . 哈耶克传 . 北京：中信出版社，2014-04-01.
[15] 张燕，张樟德 . 最实用的 120 种财务分析工具 [M]，北京：中国经济出版社，2013.
[16] 李杰 . 工业大数据 – 工业 4. 0 时代的工业转型与价值创造 [M]. 北京：机械工业出版社，2015.